A re-examination of text categorization methods

Barış Geçer
M. Ozan Karsavuran

Problem Definition

Statistical significance test on five text categorization methods with a skewed category distribution:

- SVM
- kNN
- NNet
- LLSF (Linear Least Squares Fit)
- NB

Motivation

- Cross method comparison (NNet vs SVM ?)
- Robustness on skewed category distribution
  - In real life, they are extremely non-uniform
- Effectiveness of each method as a function of rareness of categories
  - Single score : accuracy, error rate, F1 measure
    - can be dominated by common classes
  - Multi score : Micro-averaging , macro-averaging

Contributions

- Comparison of five methods on the new benchmark corpus
- Variety of statistical significance analysis and suggestion to combine them
- Performances as a function of category frequency
  - i.e. skewed category distribution

Benchmark Corpus

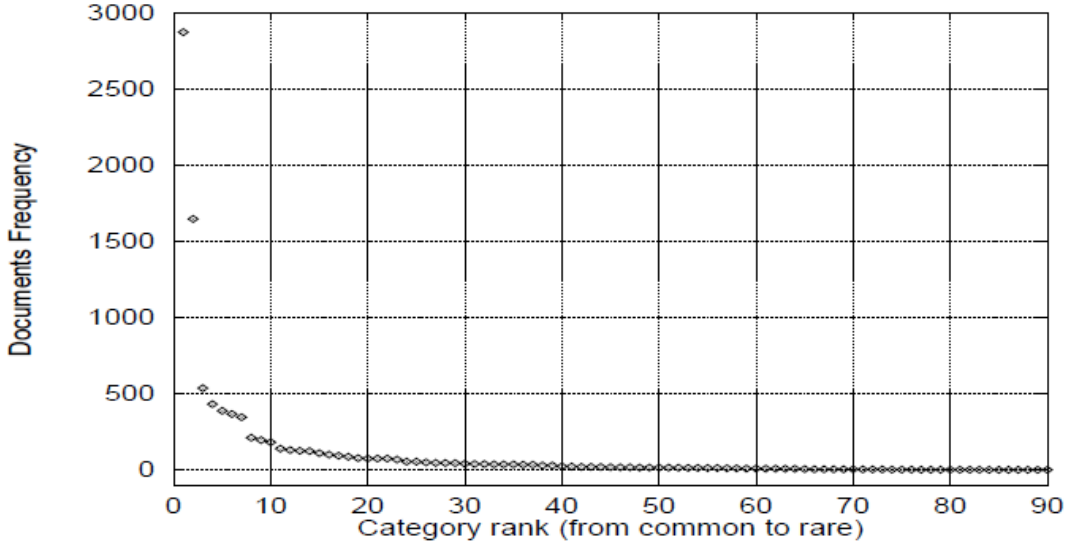

Figure 1: Category distribution in Reuters-21578 AptoMod

- %82 of categories have less than 100 instances
- %33 have less than 10

Performance Measures

- Macro-averaging : F1 measure computed for each category individually then averaged
- Micro-averaging : F1 measure computed globally
- Providing both kinds of scores is more informative than providing either alone
- Error

SVM

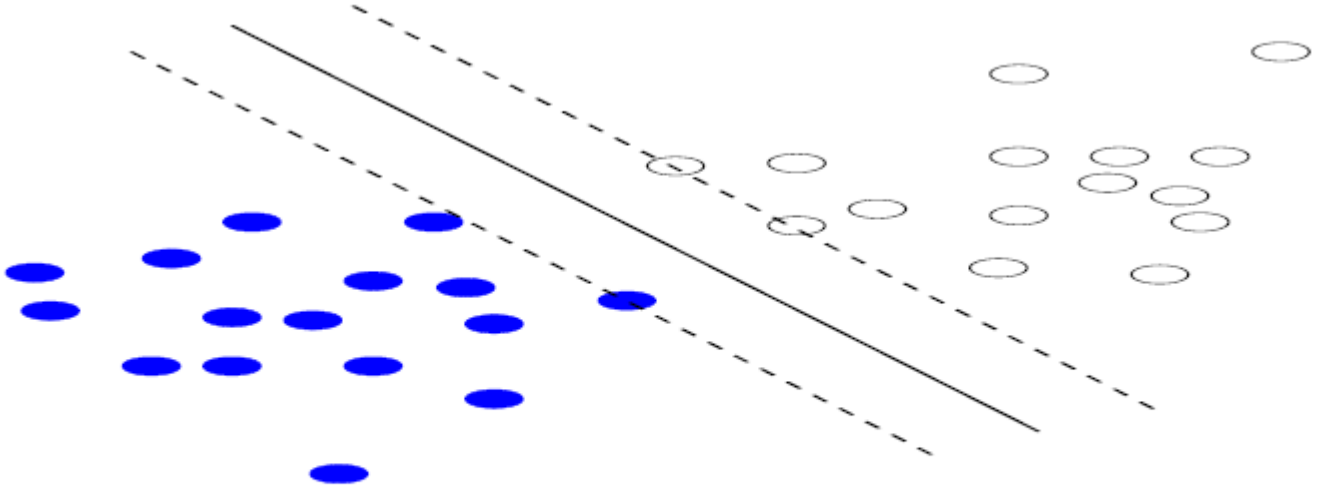

Figure 3: The decision line with the maximal margin. The data points on the dashed lines are the Support Vectors.

LLSF and kNN

- Although they differ statistically, they had similar performance in the authors' previous studies
- Yet, their robustness in dealing with rare categories is unknown.

Neural Network (NNet)

- Different Networks
- Separate NNet per category
- Training cost is high
  - One NNet for all 90 categories
  - one hidden layer

Naive Bayes (NB)

- Use joint probabilities of words and categories
  - assume words are independent

Significance Tests

- s-test and p-test at micro level
- others at macro level

- Micro sign test (s-test)
- Macro sign test (S-test)
- Macro t-test (T-test)
- Macro t-test after rank transformation
- Comparing proportions (p-test)

Significance Tests

- S-test: robust for reducing the influence of outliers but risks being insensitive
- T-test: could be overly sensitive when F scores are unstable
- T’-test: less sensitive to outliers but more sensitive than sign tests

Significance Tests

- None of them is “perfect”
  - for skewed category distribution
- So use them jointly

Evaluation

Different size of features that optimize the F score for each classifier

Comparison

Table 1: Performance summary of classifiers

method	miR	miP	miF1	maF1	error
SVM	.8120	.9137	.8599	.5251	.00365
KNN	.8339	.8807	.8567	.5242	.00385
LSF	.8507	.8489	.8498	.5008	.00414
NNet	.7842	.8785	.8287	.3765	.00447
NB	.7688	.8245	.7956	.3886	.00544

miR = micro-avg recall; miP = micro-avg prec.;
miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of SVM is lower than Joachims

but not significant

Comparison

Table 1: Performance summary of classifiers

method	miR	miP	miF1	maF1	error
SVM	.8120	.9137	.8599	.5251	.00365
KNN	.8339	.8807	.8567	.5242	.00385
LSF	.8507	.8489	.8498	.5008	.00414
NNet	.7842	.8785	.8287	.3765	.00447
NB	.7688	.8245	.7956	.3886	.00544

miR = micro-avg recall; miP = micro-avg prec.;
miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of kNN is higher than Joachims, simplified kNN is similar:

it is neither optimal nor necessary

Comparison

Table 1: Performance summary of classifiers

method	miR	miP	miF1	maF1	error
SVM	.8120	.9137	.8599	.5251	.00365
KNN	.8339	.8807	.8567	.5242	.00385
LSF	.8507	.8489	.8498	.5008	.00414
NNet	.7842	.8785	.8287	.3765	.00447
NB	.7688	.8245	.7956	.3886	.00544

miR = micro-avg recall; miP = micro-avg prec.;
miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of NB is higher,

multinomial mixture vs multivariate Bernoulli

Comparison

Table 2: Statistical significance test results

sysA	sysB	s-test	S-test	T-test	T'-test
SVM	kNN	$>$	$\sim$	$\sim$	$\sim$
SVM	LLSF	$\gg$	$\sim$	$\sim$	$\sim$
kNN	LLSF	$\gg$	$\sim$	$\sim$	$\sim$
SVM	NNet	$\gg$	$\gg$	$\gg$	$\gg$
kNN	NNet	$\gg$	$\gg$	$\gg$	$\gg$
LLSF	NNet	$\sim$	$\gg$	$\gg$	$\gg$
NB	kNN	$\ll$	$\ll$	$\ll$	$\ll$
NB	LLSF	$\ll$	$\ll$	$\ll$	$\ll$
NB	SVM	$\ll$	$\ll$	$\ll$	$\ll$
NB	NNet	$\ll$	$\sim$	$\sim$	$\sim$

Comparison

- micro level:
  - SVM > kNN >> {LLSF, NNet} >> NB
- macro level:
  - {SVM,kNN,LLSF} >> {NB,NNet}
- micro: dominated by common categories
- macro: dominated by rare categories
  - complementary

Conclusions

- Significance analysis on five well-known classifiers
- micro-level, macro-level and joint for cross comparison
- significance depends on performance measure

Thank you!

Questions & Answers