CS 533 Information Retrieval Paper Presentation

“A re-examination of text categorization methods”

Yimming Yang and Xin Liu

Semih Sahin, Abdurrahman Yasar, Tolga Yilmaz

Introduction

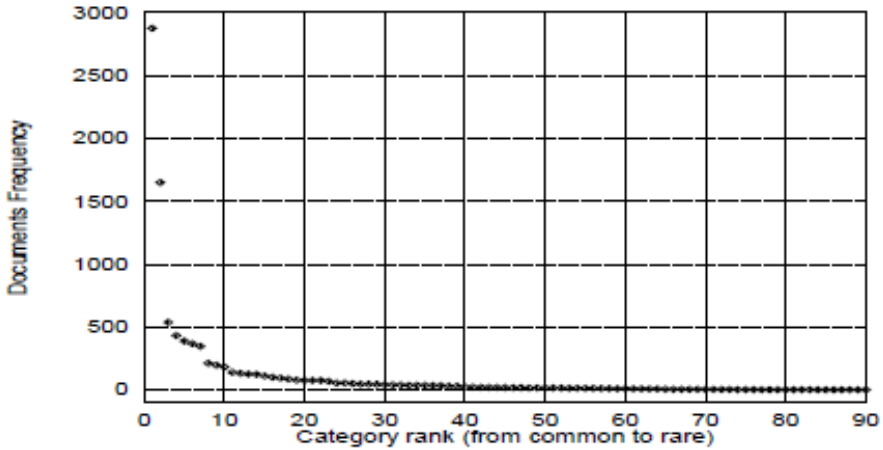
- Automatic text categorization (TC)
  - A supervised learning task : assign category labels based on previously labelled documents.
- Approaches
  - Regression models
  - Nearest Neighbor(NN) classification
  - Bayesian approaches
  - Inductive rule learning
  - Neural networks
  - On-line learning
  - Support Vector Machines (SVMs).

Introduction

- Performance of different methods are not comparable due to different data collections used in each method.
- The relation between the category distribution and the performance of methods is not fairly analyze
- This paper compares NNet,SVM,NB,kNN,LLSF
  - Reuters-21578 corpus
  - Performance of each classifier as a function of training caregory-frequency and robustness with skewed category distribution

Task and Corpus

- Task: Topic spotting of newswire stories
- Corpus : Reuters-21578
  - 7769 training documents, 3019 test documents
  - ~1.3 categories/document



Performance Measures

•Standart recall

correct assignments/ total correct assignments

• measure

(Van Rijsbergen)

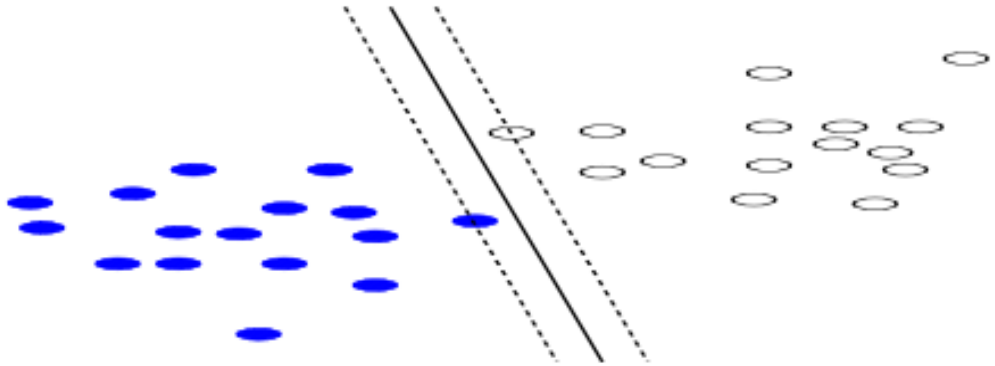
•Error

•Macro averaging : Compute for each category first, then take the average overall.

•Micro averaging: Compute globally.

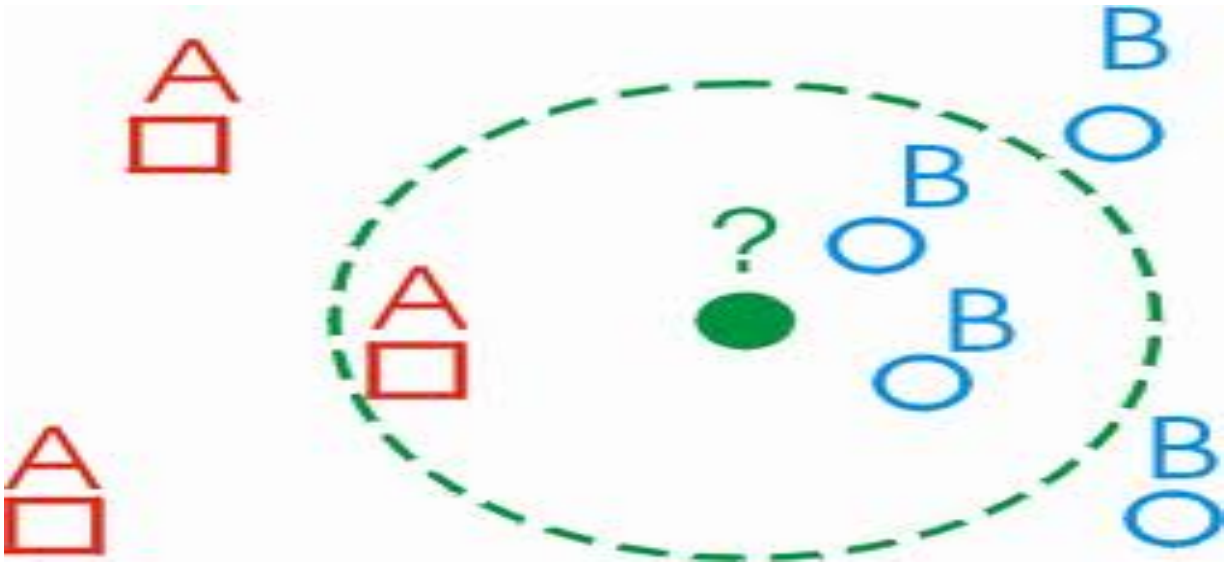
SVM(Support Vector Machines)

- Separetes vector space int c classes
- Data can be high dimensional
- Multiple SVMs can be used for non-linearly seperable data
- Example:



KNN(K Nearest Neighbors)

- Finds k nearest neighbor
- Uses them to classify the test document



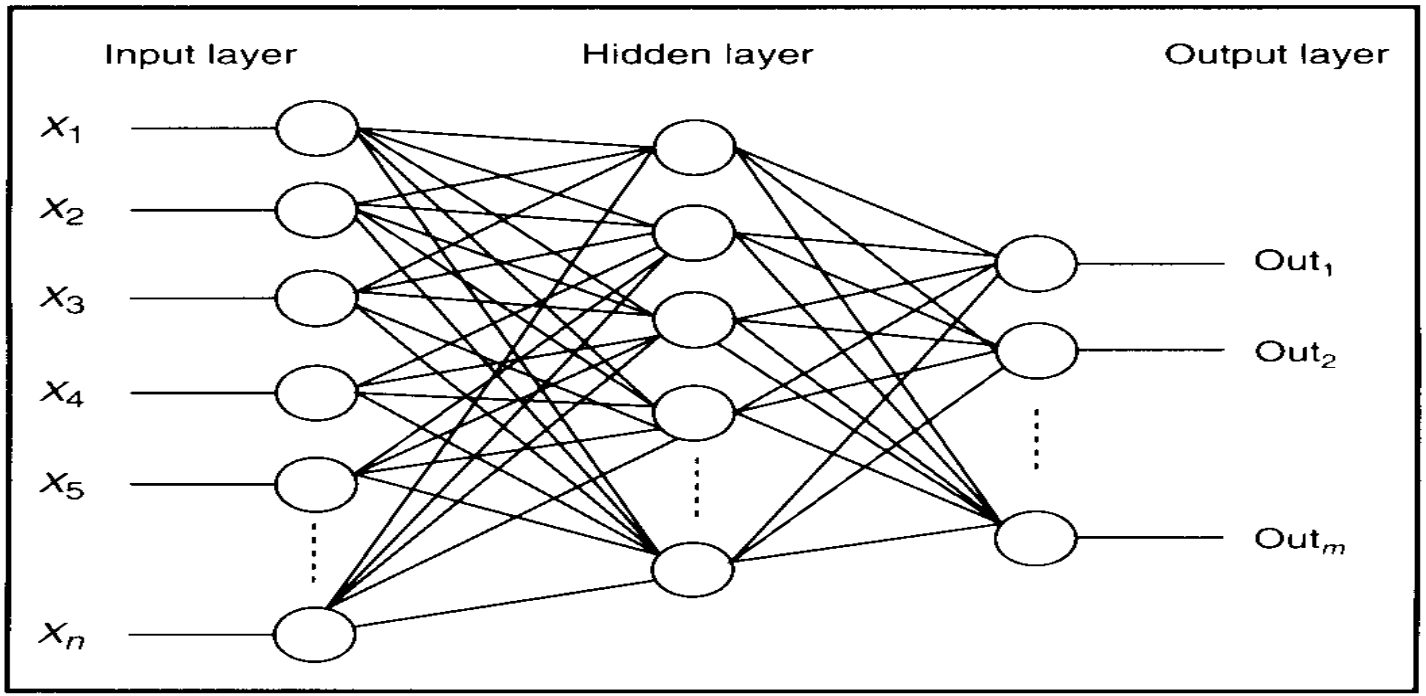
LLSF(Least Linear Square Fit

- Regression model is automatically learned from a training data
- Training data is input(document)-output(category) vectors form
- After LLSF, word-category regression coefficient matrix is obtained
- Test document is classified by its words with coefficients in FLS matrix

	Word					
Category	AIDS	and	barre	guillain	neuropathy	syndrome
acquired immunodeficiency syndrome	.198	.020	-.001	-.001	-.002	.000
nervous system diseases	-.050	.003	.020	.012	.059	.008
peripheral nerve diseases	-.003	.028	-.001	-.001	.234	.000
polyradiculoneuritis	-.001	.005	.082	.049	-.001	.032

NNet(Neural Networks)

- Interconnected group of nodes, akin to the vast network of neurons in the human brain.
- For each training document
  - System improves its learning with input word-category pair
  - This process eventually converges&learns classification



Naïve Bayes(NB)

- Consider each word independently
- This assumption makes NB computation much more efficient

- Calculate probability of each word to belong to a category
- To find category of test document for each category calculate joint probability of its words

Micro Sign Test (s-test)

- Compare 2 systems (A and B)
- Binary decisions on all documents
- Evaluates models at micro level
- Produces a (one sided) P-value for the hypothesis: A is better than B
- Smaller P -> more significant (must be <0.1)

System A	System B	Classifier	s-test[a]
BOW-NER	BOW	Forced One-vs-All	$>$
BOW-NER-GAZ	BOW	Forced One-vs-All	$\gg$
BOW-NER	BOW	Relaxed One-vs-All	$\sim$
BOW-NER-GAZ	BOW	Relaxed One-vs-All	$\sim$
BOW-NER	BOW	Multiclass	$\sim$
BOW-NER-GAZ	BOW	Multiclass	$\sim$

[a] where "$\gg$" indicates P-value ≤ 0.05; "$>$" indicates $0.05 <$ P-value ≤ 0.10; and "$\sim$" indicates P-value > 0.10

Macro Sign Test (S-test)

- Sign test designed for comparing two systems, A and B, using the paired F1 values for individual categories
- The test hyphothesis and the P-value computations are the same as those as in the micro s-test
- May be more robust for reducing the influence of outliers
- Risks being insensitive (or not sufficiently sensitive) in performance comparison because it ignores the absolute differences between F1 values

Macro t-test (T-test)

- T-test for comparing two systems, A and B,
- Use the paired F1 values for individual categories
- Notation is same as S-test
- Sensitive to the absolute values
- Could be overly sensitive when F1 scores are highly unstable (those for low-frequency categories)

Macro t-test after rank trans.

- To compare systems A and B based on the F1 values after rank transformation
- Compromise between the two extremes
- Less sensitive than T-test to outliers
- More sensitive than the sign test because it reserve the order of distinct F1 values

Comparing Proportions

- For the performance measures
  - proportions (recall, precision, accuracy or error),
- Compare the performance scores of systems
- Designed to evaluate the performance of systems at a micro level (based on the pooled decisions on individual document/ category pairs)

- recall: number of true YESes for categories
- precision: number assigned YESes by the system
- accuracy or error: number of document-category pairs

Evaluation

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.

- Micro-level analysis

SVM > kNN >> {LLSF,NNet} >>NB

- Macro-level analysis

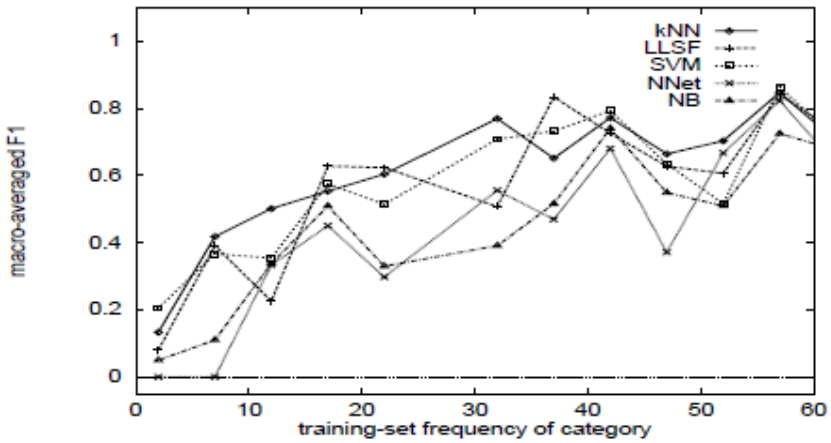
{SVM , kNN ,LLSF}>>{NB, NNet}

- Error-rate

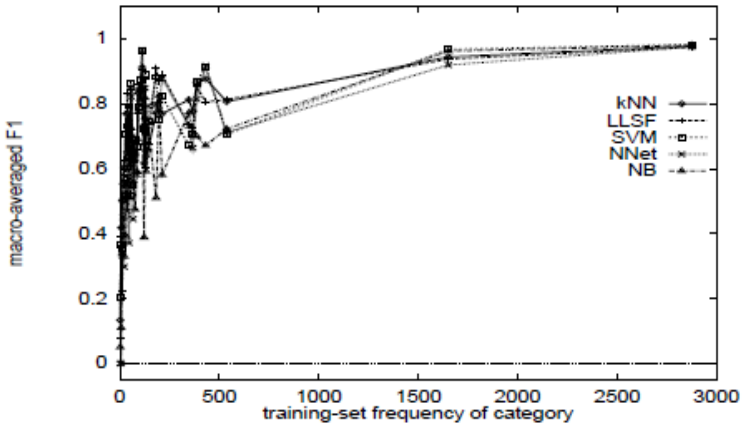
{SVM , kNN}>LLSF>NNet>>NB

Evaluation

- NNet and NB are clearly worse. The others behave similar to each other.



Performance curves on rare categories.



Performance curves on all the categories.

Conclusion

- This paper compares 5 classifiers.
- For micro-level, SVM and kNN outperform others and NB performs poorly.
- Macro-level analysis shows that SVM, kNN and LLSF behave similarly whereas NNet and NB performs significantly worse.

Questions&Remarks

THANK YOU FOR LISTENING