"A Re-examination Categorization Methods"

Overview

Reports a controlled study with statistical significance tests on five text categorization methods; SVM, kNN, NNet, LLSF, NB.

of

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

Classifiers

- 1. SVM: SVM problem can be solved using quadratic programming techniques. Algorithms for solving linearly separable cases can be extended for solving linearly non-separable cases by either.
- 2. KNN: Given a test document. Finds the k nearest neighbors among the training documents.
- 3. LLSF: Multivariate regression model is automatically learned from a training set of documents and their categories. Training data are represented in the form of input/output vector pairs. By solving a linear least-squares fit on the training pairs of vectors, one can obtain a matrix of word-category regression coefficients.
- 4. NNet: Three layered neural network with a hidden layer. Use a hidden layer with k nodes. k is empirically chosen.
- 5. NB: Use the joint probabilities of words and categories to estimate the probabilities of categories given a document. Assumes word independence. Conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category.

Significance Tests

- s-test and p-test designed to evaluate the performance of systems at a micro level based on pooled decisions on individual document/category pairs
- S-test, T-test designed to evaluate at a macro level using the performance scores on each category as the unit measure. T-test is sensitive to absolute values
- S-test is more robust for reducing the influence of outliers
- Using them together is a better idea because none of them is perfect for all the performance measures

Evaluation & Conclusion

- 1. For micro-level, SVM and kNN outperform others and NB performs poorly.
- 2. Macro-level analysis shows that SVM, kNN and LLSF behave similarly whereas NNet and NB performs significantly worse.

CS533-8 Min Presentations

[Handout]

Text

Sahin S., Yasar A., Yilmaz T.

Environment Set-up

- They applied statistical feature selection at a preprocessing stage for each classifier
- Different feature-set sizes were tested
- kNN k is set to 45. LLSF singular value set to 500. NNet – no. of hidden units is 64

Experimental Results:



Performance curves on rare categories.



Performance curves on all the categories.

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.		