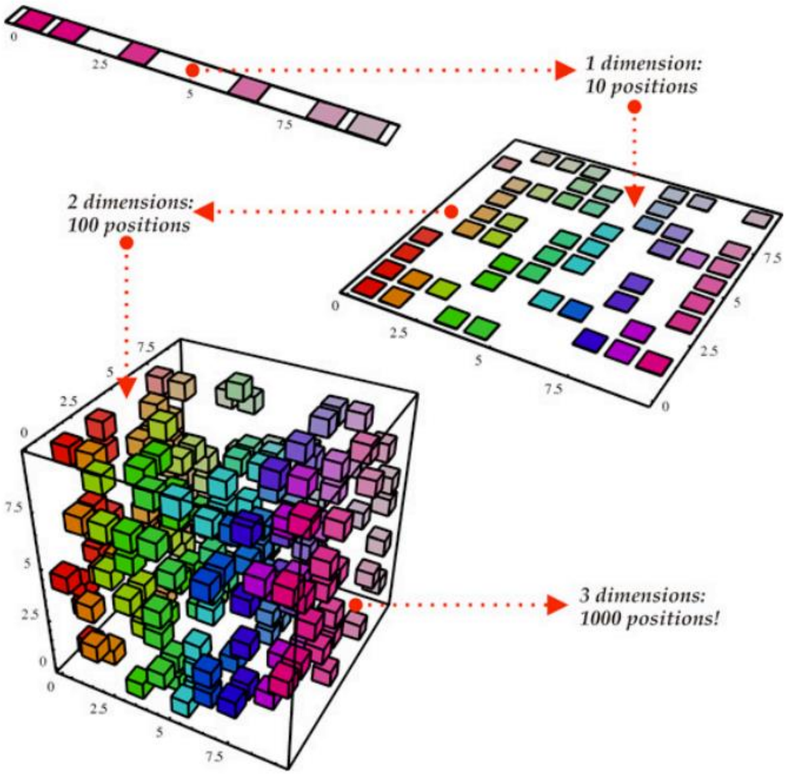


Lecture 6: Dimensionality Reduction

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Dimensionality Reduction

02 Feature Selection

03 Feature Extraction: LSA & t-SNE

Dimensionality Reduction

- Common features of text data
  - ✓ In general, a document consists of a large number of terms (words)
  - ✓ Only a few of them are actually relevant to text mining tasks even after some preprocessing (stop-words removal, stemming, lemmatization, etc)

Term Variables	Documents			
Term 1	Document1 1	Document2	···	Document n
Term 2	Data			
⋮				
Term m				

↓

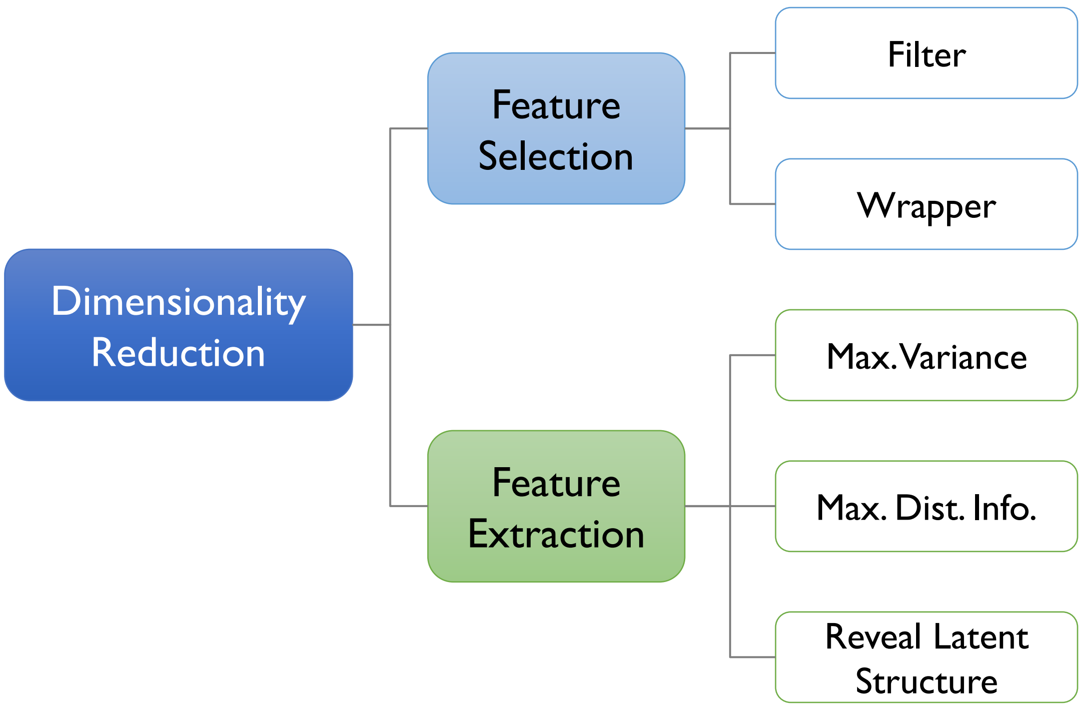
- **Problem 1**: High dimensionality (N. terms >> N. documents)
- **Problem 2**: Sparseness (Most elements in a term-document matrix are zero)

Dimensionality Reduction

- Why is dimensionality reduction necessary?
  - ✓ To make large problems computationally efficient (conserving computation, storage and network resources)
  - ✓ To improve the quality of text mining results
    - Improve classification accuracy or clustering modularity
    - Reduce the amount of training data needed to obtain a desired level of performance

Dimensionality Reduction

- A simplified taxonomy of dimensionality reduction techniques



	Class Label	Learning Algorithm
• Information Gain • Odds ratio, etc.	Used	Not used
• Forward/Backward/Stepwise • Genetic algorithm	Used	Used
• Principal component analysis	Not used	Not used
• Multidimensional scaling	Not used	Not used
• Latent semantic indexing	Not used	Not used

Dimensionality Reduction

- Feature selection vs. feature extraction
  - ✓ Feature selection: select a small subset of original variables
  - ✓ Feature extraction: construct/extract a new set of features based on the original variables

V_1	V_2	V_3	...	V_n
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Feature Selection

V_1	V_5	V_8
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

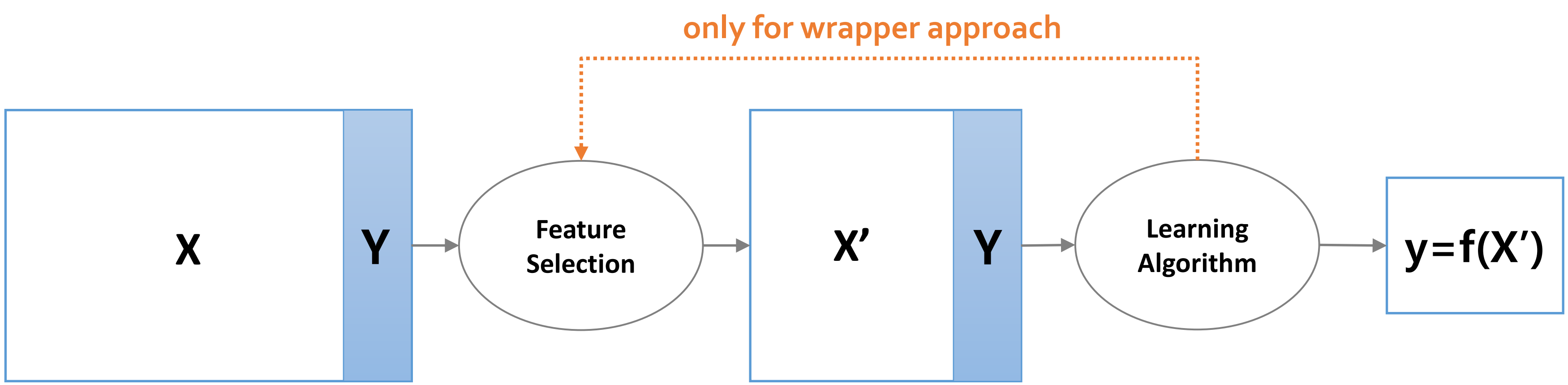
Feature Extraction

Z_1	Z_2	Z_3
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

$$Z_1 = V_1 + 0.2*V_2$$
$$Z_2 = V_3 - 2*V_5$$
$$Z_3 = V_4 + V_6 - V_9$$

Dimensionality Reduction

- Filter approach vs. Wrapper approach
  - ✓ **Filter**: select a set of features based on pre-defined criteria
    - no feedback loop, independent of the learning algorithm
  - ✓ **Wrapper**: evaluate a subset with a learning algorithm and repeat the process until a certain level of performance is achieved
    - Feedback loop exists, dependent on the learning algorithm



AGENDA

Artificial Data Set

- 10 Documents with 10 Terms
  - ✓ Binary classification/categorization problem
  - ✓ 6 positive documents & 4 negative documents
  - ✓ Binary Term-Document matrix

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1
Term 4	1	1	1	1	1	1	1	1	0	0
Term 5	0	0	0	1	1	1	1	1	1	1
Term 6	1	1	1	0	0	0	0	0	0	0
Term 7	0	0	0	0	0	0	1	1	0	0
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0
Term 10	1	0	0	0	0	0	0	0	1	1
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg

Feature Selection Metric 1-4

- Document frequency (DF)

  ✓ Simply count the number of total documents in which a word *w* is presented

$$DF(w) = N_D(w)$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1

- For Term 1: DF(w) = 6
- For Term 2: DF(w) = 4
- For Term 3: DF(w) = 10

Feature Selection Metric 1-4

- Accuracy (Acc)
  - ✓ Expected accuracy of a simple classifier built from the single feature

$$Acc(w) = N(Pos, w) - N(Neg, w)$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1

- For Term 1: $N(Pos, w) = 6, N(Neg, w) = 0,\ Acc(w) = 6$
- For Term 2: $N(Pos, w) = 0, N(Neg, w) = 4,\ Acc(w) = -4$
- For Term 3: $N(Pos, w) = 6, N(Neg, w) = 4,\ Acc(w) = 2$

Feature Selection Metric 1-4

- Accuracy ratio (AccR)
  - ✓ Expected accuracy of a simple classifier built from the single feature

$$AccR(w) = \left| \frac{N(Pos,w)}{N(Pos)} - \frac{N(Neg,w)}{N(Neg)} \right|$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1

- For Term 1: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{0}{4} = 0, AccR(w) = 1$
- For Term 2: $\frac{N(Pos,w)}{N(Pos)} = \frac{0}{6} = 0, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 1$
- For Term 3: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 0$

Feature Selection Metric 1-4

- Probability Ratio (PR)
  - ✓ The probability of the word given the positive class divided by the probability of the word given the negative class

$$PR(w) = \frac{N(Pos,w)}{N(Pos)} / \frac{N(Neg,w)}{N(Neg)}$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1

- For Term 1: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{0}{4} = 0, PR(w) = \infty$
- For Term 2: $\frac{N(Pos,w)}{N(Pos)} = \frac{0}{6} = 0, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 0$
- For Term 3: $\frac{N(Pos,w)}{N(Pos)} = \frac{6}{6} = 1, \frac{N(Neg,w)}{N(Neg)} = \frac{4}{4} = 1, AccR(w) = 1$

Feature Selection Metric 1-4

- Compute the metric 1-4 for the data set

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	DF	Acc	AccR	PR
Term 1	1	1	1	1	1	1	0	0	0	0	6	6	1.00	Inf
Term 2	0	0	0	0	0	0	1	1	1	1	4	-4	1.00	0.00
Term 3	1	1	1	1	1	1	1	1	1	1	10	2	0.00	1.00
Term 4	1	1	1	1	1	1	1	1	0	0	8	4	0.50	2.00
Term 5	0	0	0	1	1	1	1	1	1	1	7	-1	0.50	0.50
Term 6	1	1	1	0	0	0	0	0	0	0	3	3	0.50	Inf
Term 7	0	0	0	0	0	0	1	1	0	0	2	-2	0.50	0.00
Term 8	1	0	1	0	1	0	1	0	1	0	5	1	0.00	1.00
Term 9	1	1	1	0	0	0	1	0	0	0	4	2	0.25	2.00
Term 10	1	0	0	0	0	0	0	0	1	1	3	-1	0.33	0.33
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg				

Feature Selection Metric 5-6

- Odds ratio (OddR)
  - ✓ Reflect the odds of the word occurring in the positive class normalized by that of the negative class
    - It has been used for relevance ranking in information retrieval

$$OddR(w) = \frac{N(Pos, w)}{N(Neg, w)} \times \frac{N(Neg, \overline{w})}{N(Pos, \overline{w})}$$

    - Add 1 to any zero count in the denominator to avoid division by zero

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0

- For Term 8: $\frac{N(Pos,w)}{N(Neg,w)} = \frac{3}{2}, \frac{N(Neg,\overline{w})}{N(Pos,\overline{w})} = \frac{2}{3}, OddR(w) = 1$
- For Term 9: $\frac{N(Pos,w)}{N(Neg,w)} = \frac{3}{1}, \frac{N(Neg,\overline{w})}{N(Pos,\overline{w})} = \frac{3}{3}, OddR(w) = 3$

Feature Selection Metric 5-6

- Odds ratio Numerator (OddN)

$$OddN(w) = N(Pos, w) \times N(Neg, \overline{w})$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 8	1	0	1	0	1	0	1	0	1	0
Term 9	1	1	1	0	0	0	1	0	0	0

- For Term 8: $N(Pos, w) = 3$, $N(Neg, \overline{w}) = 2$, $OddN(w) = 6$
- For Term 9: $N(Pos, w) = 3$, $N(Neg, \overline{w}) = 3$, $OddN(w) = 9$

Feature Selection Metric 7

- F1-Measure
  - ✓ Expected accuracy of a simple classifier built from the single feature

$$F1(w) = \frac{2 \times Recall(w) \times Precision(w)}{Recall(w) + Precision(w)}$$

$$Recall(w) = \frac{N(Pos, w)}{N(Pos, w) + N(Pos, \overline{w})}, \qquad Precision(w) = \frac{N(Pos, w)}{N(Pos, w) + N(Neg, w)}$$

  - ✓ By doing some arithmetic operations, we can derive

$$F1(w) = \frac{2 \times N(Pos, w)}{N(Pos) + N(w)}$$

  - ✓ In F1 measure, negative features are devalued compared to positive features

Feature Selection Metric 7

- F1-Measure

$$F1(w) = \frac{2 \times N(Pos, w)}{N(Pos) + N(w)}$$

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Term 1	1	1	1	1	1	1	0	0	0	0
Term 2	0	0	0	0	0	0	1	1	1	1
Term 3	1	1	1	1	1	1	1	1	1	1

- For Term 1: $F1(w) = \frac{2 \times 6}{6+6} = 1$
- For Term 2: $F1(w) = \frac{2 \times 0}{6+4} = 0$
- For Term 3: $F1(w) = \frac{2 \times 6}{6+10} = 0.75$

Feature Selection Metric 5-7

- Compute the metric 5-7 for the data set

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	OddR	OddN	F1
Term 1	1	1	1	1	1	1	0	0	0	0	24.00	24	1.00
Term 2	0	0	0	0	0	0	1	1	1	1	0.00	0	0.00
Term 3	1	1	1	1	1	1	1	1	1	1	0.00	0	0.75
Term 4	1	1	1	1	1	1	1	1	0	0	4.00	12	0.86
Term 5	0	0	0	1	1	1	1	1	1	1	0.00	0	0.46
Term 6	1	1	1	0	0	0	0	0	0	0	4.00	12	0.67
Term 7	0	0	0	0	0	0	1	1	0	0	0.00	0	0.00
Term 8	1	0	1	0	1	0	1	0	1	0	1.00	6	0.55
Term 9	1	1	1	0	0	0	1	0	0	0	3.00	9	0.60
Term 10	1	0	0	0	0	0	0	0	1	1	0.20	2	0.22
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg			

Feature Selection Metric 8

- Information Gain: IG
  - ✓ Measures the decrease in entropy when the feature is given vs. absent.
  - ✓ Entropy without the information provided by the term w

$$Entropy(absent\ w) = \sum_{C \in \{Pos,Neg\}} -P(C) \times \log(P(C))$$

$$Entropy(given\ w) = P(w)\left[\sum_{C \in \{Pos,Neg\}} -P(C|w) \times \log(P(C|w))\right] + P(\overline{w})\left[\sum_{C \in \{Pos,Neg\}} -P(C|\overline{w}) \times \log(P(C|(\overline{w}))\right]$$

$$IG(w) = Entropy(absent\ w) - Entropy(given\ w)$$

Feature Selection Metric 8

- Information Gain: IG

  ✓ For Term 1

$$Entropy(absent\ w) = -P(Pos) \times \log(P(Pos)) - P(Neg) \times \log(P(Neg))$$
$$= -0.6 \times \log(0.6) - 0.4 \times \log(0.4)$$
$$= 0.29$$

$$Entropy(given\ w) = P(w)[-P(Pos|w) \times \log(P(Pos|w)) - P(Neg|w) \times \log(P(Neg|w))]$$
$$+ P(\overline{w})\left[-P(Pos|\overline{w}) \times \log(P(Pos|\overline{w})) - P(Neg|\overline{w}) \times \log\left(P(Neg|\overline{w})\right)\right]$$
$$= 0.6[-1 \times \log(1) - 0 \times \boxed{\log(0)}] + 0.4[-0 \times \boxed{\log(0)} - 1 \times \log(1)]$$
$$= 0$$

Convert log(o) to zero

$$IG(w) = 0.29 - 0 = 0.29$$

Feature Selection Metric 9

- Chi-squared statistic (χ^2)
  - ✓ Measures divergence from the distribution expected if one assumes the feature occurrence is independent of the class label

$$\chi^2(w) = \frac{N \times [P(Pos,w) \times P(Neg,\overline{w}) - P(Neg,w) \times P(Pos,\overline{w})]^2}{P(w) \times P(\overline{w}) \times P(Pos) \times P(Neg)}$$

Term 1	Pos	Neg	Total
w	6	0	6
$\overline{w}$	0	4	4
total	6	4	10

Term 4	Pos	Neg	Total
w	6	2	8
$\overline{w}$	0	2	2
total	6	4	10

$$\chi^2(T1) = \frac{10 \times [0.6 \times 0.4 - 0 \times 0]^2}{0.6 \times 0.4 \times 0.6 \times 0.4} = 10.00$$

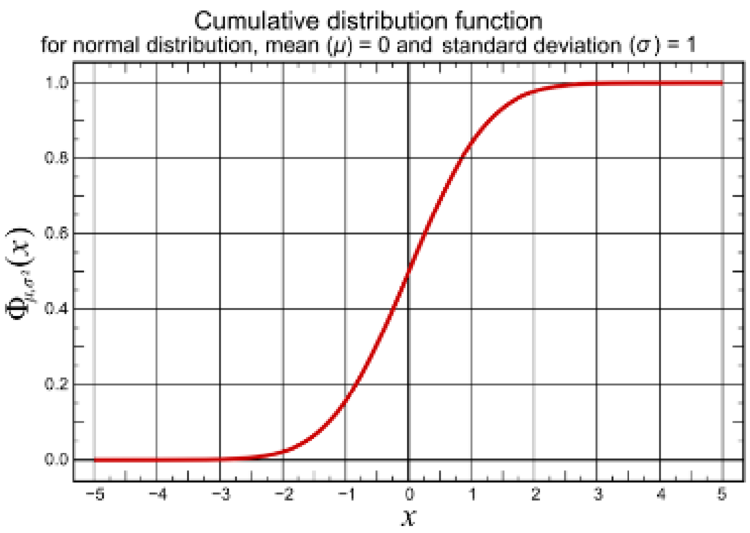
$$\chi^2(T4) = \frac{10 \times [0.6 \times 0.2 - 0.2 \times 0]^2}{0.8 \times 0.2 \times 0.6 \times 0.4} = 3.75$$

Feature Selection Metric 10

- Bi-Normal Separation (BNS)
  - ✓ Measures the degree of separation assuming that the occurrence of a feature in a document is a random process following a normal distribution

$$BNS(w) = \left| F^{-1}\left(\frac{N(Pos, w)}{N(Pos)}\right) - F^{-1}\left(\frac{N(Neg, w)}{N(Neg)}\right) \right|$$

F: c.d.f of the standard normal distribution



Term 4	Pos	Neg	Total
w	6	2	8
$\bar{w}$	0	2	2
total	6	4	10

$$BNS(w) = |F^{-1}(1) - F^{-1}(0.5)|$$

$$\approx |F^{-1}(0.9995) - F^{-1}(0.5)|$$

$$= |3.29 - 0| = 3.29$$

Feature Selection Metric 8-10

- Compute the metric 8-10 for the data set

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	IG	χ^2	BNS
Term 1	1	1	1	1	1	1	0	0	0	0	0.29	10.00	6.58
Term 2	0	0	0	0	0	0	1	1	1	1	0.29	10.00	6.58
Term 3	1	1	1	1	1	1	1	1	1	1	0.00	0.00	0.00
Term 4	1	1	1	1	1	1	1	1	0	0	0.10	3.75	3.29
Term 5	0	0	0	1	1	1	1	1	1	1	0.08	2.86	3.29
Term 6	1	1	1	0	0	0	0	0	0	0	0.08	2.86	3.29
Term 7	0	0	0	0	0	0	1	1	0	0	0.10	3.75	3.29
Term 8	1	0	1	0	1	0	1	0	1	0	0.00	0.00	0.00
Term 9	1	1	1	0	0	0	1	0	0	0	0.01	0.63	0.67
Term 10	1	0	0	0	0	0	0	0	1	1	0.03	1.27	0.97
Class	Pos	Pos	Pos	Pos	Pos	Pos	Neg	Neg	Neg	Neg			

Feature Selection Metric: Summary

- Comparison of the 10 feature selection metrics
  - ✓ For the positive class, the Term 4 is included for the top 3 variables by all metrics, followed by Term 1 and Term 6

	DF	Acc	AccR	PR	OddR	OddN	F1	IG	χ^2	BNS	Top3
Term 1	6	6	1.00	Inf	24.00	24	1.00	0.29	10.00	6.58	9
Term 2	4	-4	1.00	0.00	0.00	0	0.00	0.29	10.00	6.58	4
Term 3	10	2	0.00	1.00	0.00	0	0.75	0.00	0.00	0.00	2
Term 4	8	4	0.50	2.00	4.00	12	0.86	0.10	3.75	3.29	10
Term 5	7	-1	0.50	0.50	0.00	0	0.46	0.08	2.86	3.29	3
Term 6	3	3	0.50	Inf	4.00	12	0.67	0.08	2.86	3.29	6
Term 7	2	-2	0.50	0.00	0.00	0	0.00	0.10	3.75	3.29	4
Term 8	5	1	0.00	1.00	1.00	6	0.55	0.00	0.00	0.00	0
Term 9	4	2	0.25	2.00	3.00	9	0.60	0.01	0.63	0.67	1
Term 10	3	-1	0.33	0.33	0.20	2	0.22	0.03	1.27	0.97	0

Empirical Study

Forman (2003)

- Empirical study conducted by Forman (2003)
  - ✓ Data sets: 229 text classification tasks (from Reuters, TREC, OHSUMED, etc.)
  - ✓ SVM as a base classifier, one-against-all method for multiclass problems
  - ✓ Performances are evaluated in terms of accuracy, precision, recall, and F-1 measure

- Analysis purpose
  - ✓ To obtain the best overall classification performance regardless of the number of features
  - ✓ To find the best metric when only a very small number of features is selected
    - For limited resources, fast classification, and large scalability
  - ✓ Contract the performance under high-skew and low-skew class distribution situations

Empirical Study

Forman (2003)

- Metrics considered

Name	Description	Formula
Acc	Accuracy	$tp - fp$
Acc2	Accuracy balanced[†]	$\lvert tpr - fpr \rvert$
BNS	Bi-Normal Separation[†]	$\lvert F^{-1}(tpr) - F^{-1}(fpr) \rvert$ where F is the Normal c.d.f.
Chi	Chi-Squared[‡]	$t(tp, (tp+fp)P_{pos}) + t(fn, (fn+tn)P_{pos}) +$ $t(fp, (tp+fp)P_{neg}) + t(tn, (fn+tn)P_{neg})$ where $t(count, expect) = (count - expect)^2 / expect$
DFreq	Document Frequency[†‡°]	$tp + fp$
F1	F_1-Measure	$\frac{2\, recall\ precision}{(recall + precision)} = \frac{2\, tp}{(pos + tp + fp)}$
IG	Information Gain[†‡]	$e(pos, neg) - [\, P_{word}\, e(tp, fp) + P_{\overline{word}}\, e(fn, tn) \,]$ where $e(x, y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$
OddN	Odds Ratio Numerator	$tpr\,(1 - fpr)$
Odds	Odds Ratio[†]	$\frac{tpr\,(1-fpr)}{(1-tpr)\,fpr} = \frac{tp\ tn}{fp\ fn}$
Pow	Power	$(1 - fpr)^k - (1 - tpr)^k$ where k=5
PR	Probability Ratio	tpr / fpr
Rand	Random[‡°]	random()

† Acc2, BNS, DFreq, IG, and Odds select a substantial number of negative features.
‡ Chi, IG, DFreq, and Rand also generalize for multi-class problems.
° DFreq and Rand do not require the class labels.

Empirical Study

Forman (2003)

- Experimental result (1/5)
  - ✓ BNS performed best by a wide margin when using 500 to 1,000 features
  - ✓ IG can achieve slightly better performance then the model with all features

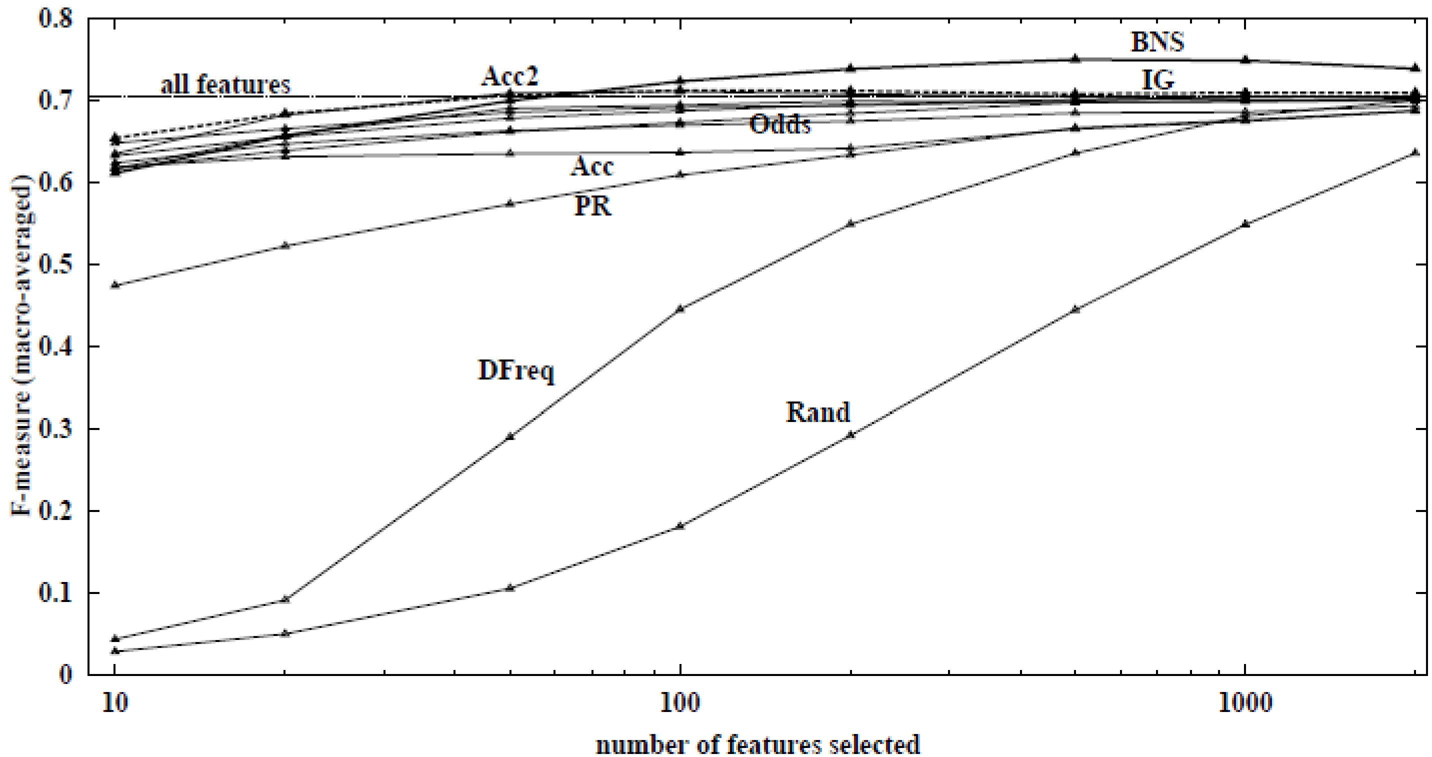


Figure 4. F-measure averaged over 229 problems for each metric, varying the number of features.

Empirical Study

Forman (2003)

- Experimental result (2/5)
  - ✓ A high recall of BNS contributes to a high F1-measure compared to others
  - ✓ If precision is the central goal, IG and χ^2 can be good choices

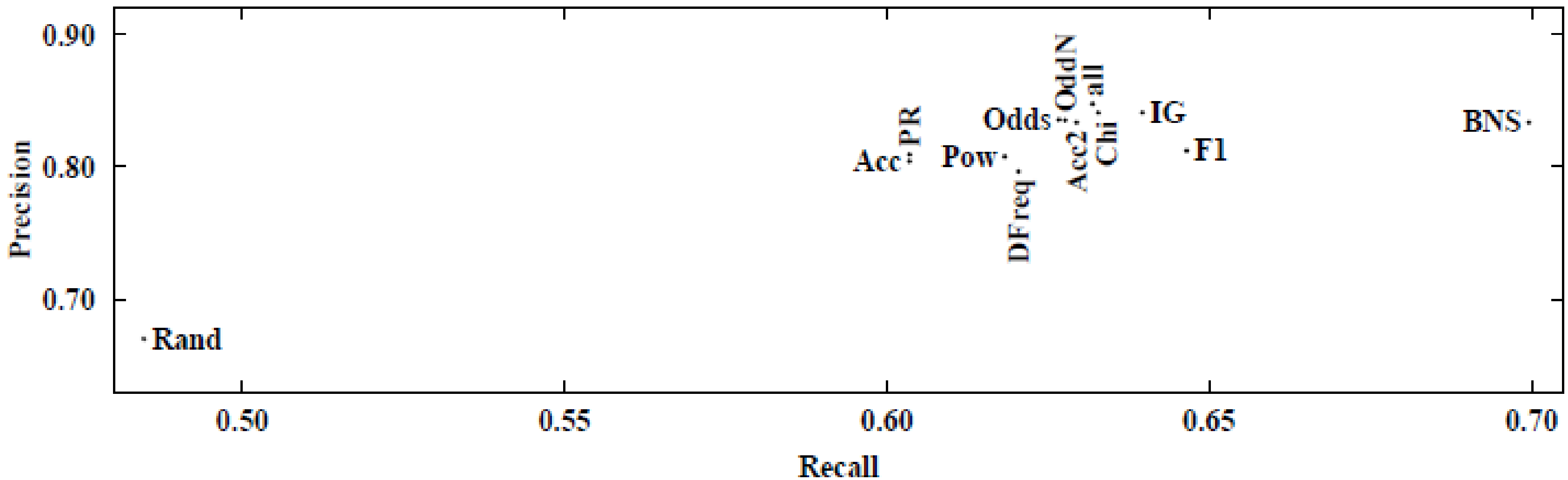


Figure 5. Precision-Recall tradeoffs from Figure 4 at 1000 features selected.

Empirical Study

Forman (2003)

- Experimental result (3/5)
  - ✓ Performances are degraded when the degree of class imbalance increases

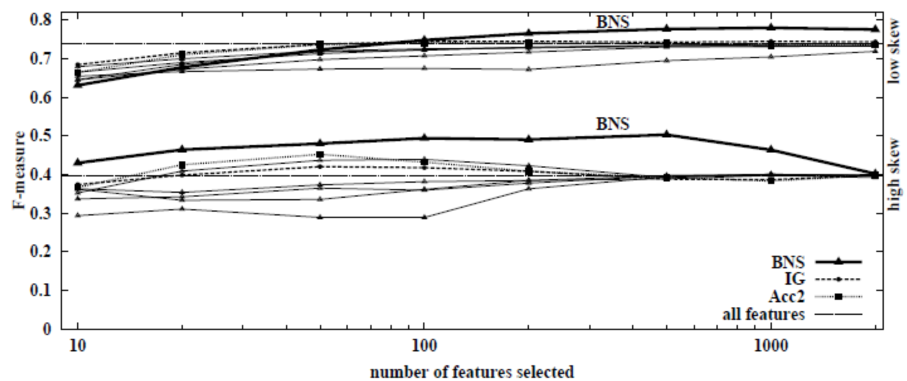


Figure 7. Average F-measure for each metric in low-skew and high-skew situations (threshold 1:67, the 90th percentile), as we vary the number of features. (To improve readability, we omitted Rand, DFreq, and PR.)

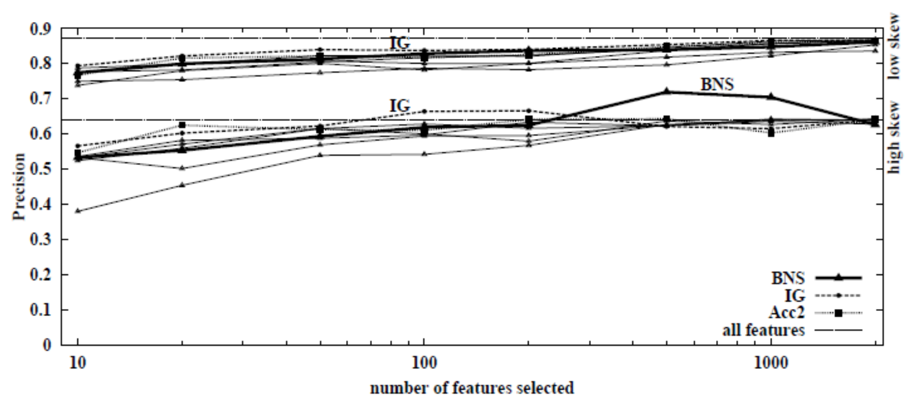


Figure 8. As Figure 7, but for precision.

Empirical Study

Forman (2003)

- Experimental result (4/5)
  - ✓ In terms of F-measure, BNS performed better than other feature selection metrics, followed by IG, and χ^2

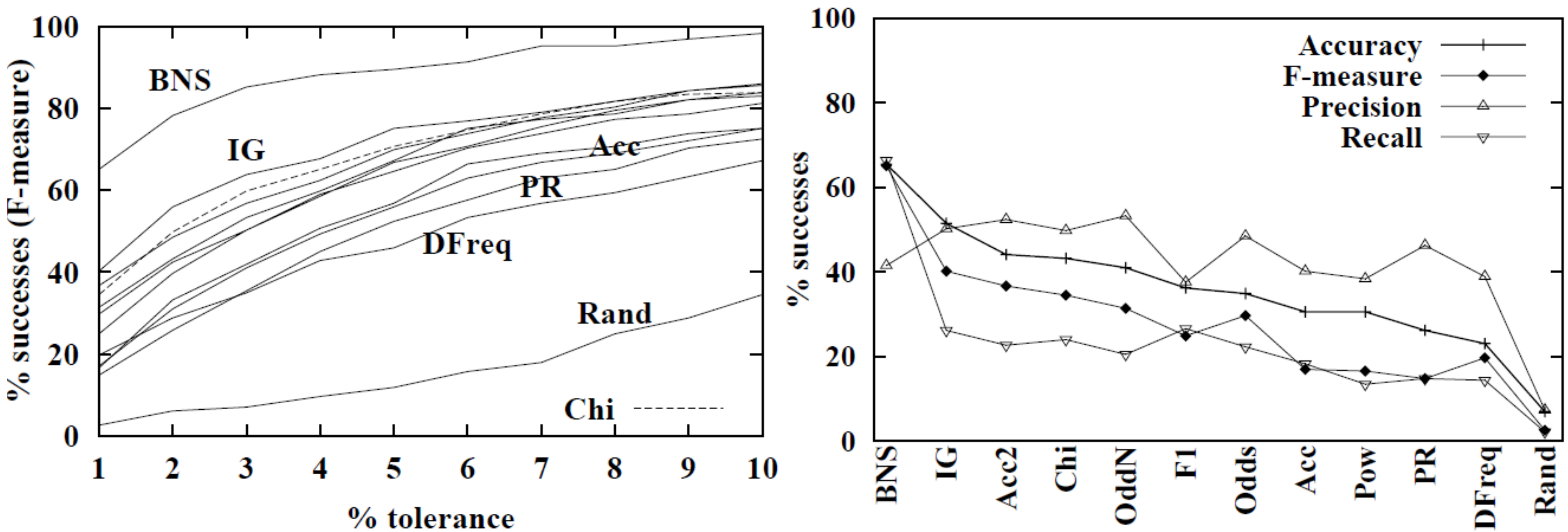


Figure 9. (a) Percentage of problems on which each metric scored within x% tolerance of the best F-measure of any metric. (b) Same, for F-measure, recall, and precision at a fixed tolerance of 1%, and for accuracy at a tolerance of 0.1%.

Empirical Study

Forman (2003)

- Experimental result (5/5)
  - ✓ In terms of precision, IG performed better than other metrics

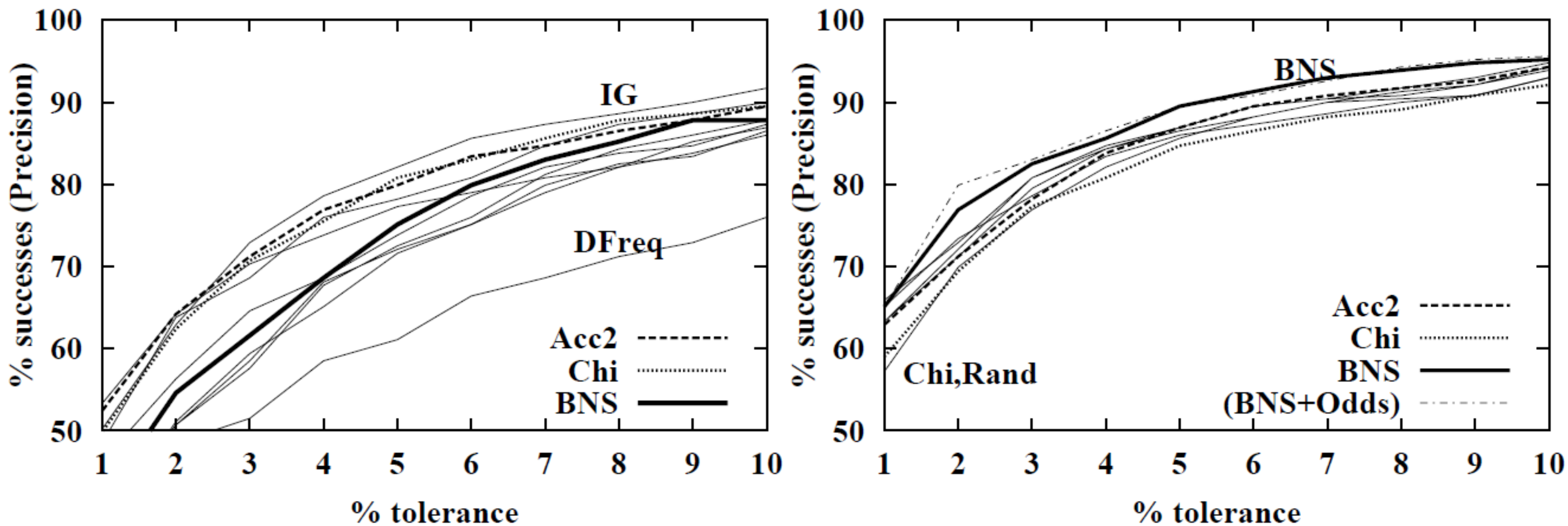


Figure 10. (a) As Figure 9a, but for precision. (b) Same axes and scale, but for each metric combined with IG. (Except the BNS+Odds curve is not combined with IG.)


ANY
questions?