O testování shody ROC křivek

J. Antoch, M. Betinec, L. Prchal a P. Sarda

Univerzita Karlova v Praze
Univerzité Paul Sabatier, Toulouse

4. února 2010

REAL PROBLEM

Real problem – collocation extraction

Problem

Our colleagues from the Institute of Formal and Applied Linguistics (ÚFAL) are developing automatic method for two-word collocation extraction from a text corpus PDT 2.0 comprising more than 2.10^6 annotated sentences.

Examples of bigram collocations

- visí otazník – the question mark is hanging – open question
- mít pravdu – to have right – to be right

Data set available

- 2 557 have been annotated as true collocations . . . $\mathcal{C}_1$
- 9 675 have been annotated as normal bigrams . . . $\mathcal{C}_0$

Stochastic point of view

Classification approach

Candidate's chance to be a true collocation is evaluated using a so called association measure $Y \equiv Y(g, \vartheta)$. ($\Rightarrow$ hyperparameter)
These measures are supposed to separate $\mathcal{C}_0$ and $\mathcal{C}_1$ linearly

$$Y \geq \vartheta \Longrightarrow g \in \mathcal{C}_1$$
$$Y < \vartheta \Longrightarrow g \in \mathcal{C}_0$$

for a collocation candidate g and an arbitrary threshold $\vartheta \in \mathbb{R}$.

Decision problem

The problem can be seen as a statistical decision $g \in \mathcal{C}_0$ against $g \in \mathcal{C}_1$ with a "critical value" ϑ

General aim

To measure/display overall performance of Y for varying ϑ

Analogy with statistical testing

Choice of ϑ

$$Y \geq \vartheta \Longrightarrow g \in \mathcal{C}_1$$
$$Y < \vartheta \Longrightarrow g \in \mathcal{C}_0$$

Small values of ϑ

- $\mathcal{C}_1$ is “preferred”
- the primary interest in “high power”
- almost all true collocations are labeled with a big amount of wrongly labeled normal words – very fine translation

Large values of ϑ

- the primary interest in “small level”
- only the most evident true collocations are labeled with a small amount of misclassified normal words – rough translation

ROC REMINDER

Two-sample classification

Objects

$\mathcal{C}$ . . . set of objects

- $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1, \quad \mathcal{C}_0 \cap \mathcal{C}_1 = \emptyset$
- $\mathcal{C}_0$ . . . class without condition
- $\mathcal{C}_1$ . . . class with condition

Condition

Existence of considered event

- illness
- bonita etc.

Reality

For $i = 1, \ldots, n$

$$i\text{th object} \in \begin{cases} \mathcal{C}_0 \\ \mathcal{C}_1 \end{cases} \equiv G_i = \begin{cases} 0 \\ 1 \end{cases}$$

Linear classifier

Diagnostic variable Y

- (discriminant) score, marker etc.
- evaluation of object properties

Treshold value ϑ and decision

$$\widehat{G}(\vartheta) = \begin{cases} 0 & \text{if } Y \leq \vartheta \\ 1 & \text{if } Y > \vartheta, \end{cases} \quad \vartheta \in \mathbb{R}$$

Hyperparameter

- given by classification method used up to the value of $\Rightarrow$ hyperparameter $\boldsymbol{\alpha} \in \mathcal{A}$

$$Y : \mathfrak{X} \times \mathcal{A} \longrightarrow \mathbb{R}$$
$$(\boldsymbol{x}, \boldsymbol{\alpha})^T \longmapsto y$$

- covers most typical methods as LDA, QDA, FLDA, LogReg, NNet, SVM, ...

Evaluation of classifier

Representation of classifier

- diagnostic variable Y
- fixed choice of hyperparameter α ... training

Description of behavior

- for one $\vartheta \Rightarrow$ one classification ... $\Rightarrow$ traditional criteria (Acc, Err, risk)
- for all $\vartheta \in \mathbb{R} \Rightarrow$ all possible classifications ... ROC curve

ROC curve definition

RANGE OF VALUES

$$\begin{aligned} \boldsymbol{r}: \ \mathbb{R} &\longrightarrow [0,1]\times[0,1] \\ \vartheta &\longmapsto [\mathrm{FPR}(\vartheta),\ \mathrm{TPR}(\vartheta)] = [1-F_0(\vartheta), 1-F_1(\vartheta)] \end{aligned}$$

True Positive Rate .... sensitivity, recall, hit rate

$$\mathrm{TPR}(\vartheta) = \mathrm{P}\left(\widehat{G}(Y,\vartheta) = 1 \mid G = 1\right) = 1 - F_1(\vartheta)$$

False Positive Rate .... nonspecificity, fallout, alarm rate

$$\mathrm{FPR}(\vartheta) = \mathrm{P}\left(\widehat{G}(Y,\vartheta) = 1 \mid G = 0\right) = 1 - F_0(\vartheta)$$

$$\boxed{F_0(y) = \mathrm{P}(Y \le y \mid G = 0)} \quad \text{and} \quad \boxed{F_1(y) = \mathrm{P}(Y \le y \mid G = 1)}$$

ROC curve definition

Theoretical ROC curve is the range of

$$\varrho(\,\cdot\,; F_0, F_1): \; \mathbb{R} \to [0,1] \times [0,1]$$
$$\vartheta \mapsto \left[1 - F_0(\vartheta), 1 - F_1(\vartheta)\right].$$

where

$$F_0(\vartheta) = \mathsf{P}(Y \leq \vartheta \,|\, \mathcal{C}_0) \equiv \mathsf{P}(Y_0 \leq \vartheta),$$
$$F_1(\vartheta) = \mathsf{P}(Y \leq \vartheta \,|\, \mathcal{C}_1) \equiv \mathsf{P}(Y_1 \leq \vartheta)$$

It is a curve in $[0,1] \times [0,1]$ square consisting of $1 - F_1(\vartheta)$ on the vertical axis plotted against $1 - F_0(\vartheta)$ on the horizontal axis $\forall t \in \mathbb{R}$

$$\boxed{\mathrm{ROC}_Y = \left\{ \boldsymbol{r} \in [0,1]^2 : \exists \vartheta \in \mathbb{R} \quad \varrho(\vartheta; F_0, F_1) = \boldsymbol{r} \right\}}$$

Collocation extraction revisited

Aim

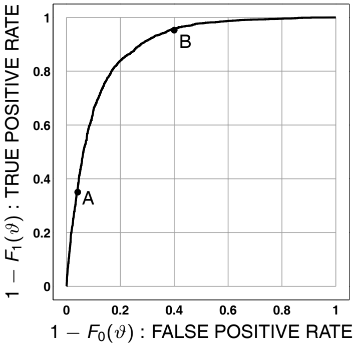
To measure/display overall performance of Y for varying ϑ.

Rough translation

A . . . most evident collocations are labeled; level 5 %, power 40 %

Fine translation

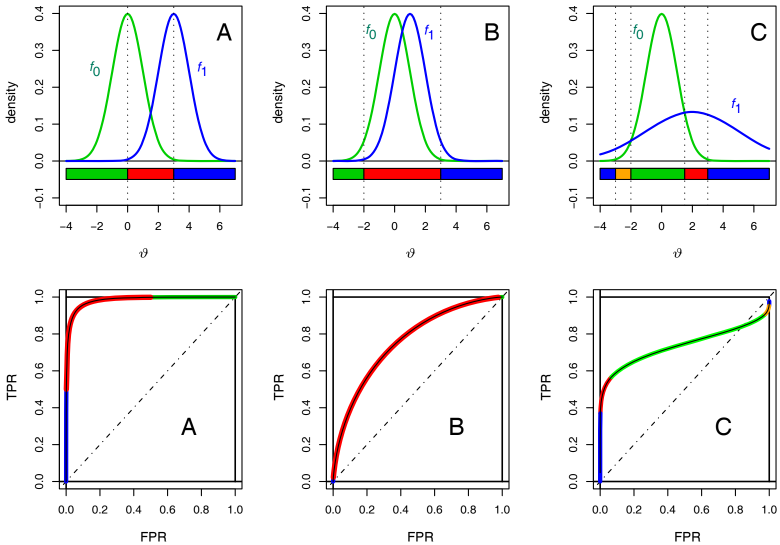
B . . . almost all collocations are labeled; level 40 %, power 95 %



Remind

$$F_0(\vartheta) = \mathrm{P}(Y \leq \vartheta \,|\, \mathcal{C}_0) \equiv \mathrm{P}(Y_0 \leq \vartheta) \;\;\&\;\; F_1(\vartheta) = \mathrm{P}(Y \leq \vartheta \,|\, \mathcal{C}_1) \equiv \mathrm{P}(Y_1 \leq \vartheta)$$

Examples of ROC curves



Alternative definition

ROC curve – TPR as function of FPR

$$\mathrm{ROC}(\xi) = \mathrm{TPR}(\xi) = 1 - F_1\big(F_0^{-1}(1-\xi)\big),$$
$$\text{where} \quad \xi := \mathrm{FPR} \in [0,1]$$

Properties

- Assumptions
  - F_0 and F_1 are absolutely continuous (may be weakened)
  - supports f_0 and f_1 are identical
- used in parametric models

Remind

$$\begin{aligned} TPR &= 1 - F_1\big(F_0^{-1}(1-FPR)\big) = 1 - F_1\big(F_0^{-1}(1-(1-F_0(\vartheta)))\big) \\ &= 1 - F_1(\vartheta) \equiv TPR(\vartheta) \end{aligned}$$

NONPARAMETRIC APPROACH

ROC curves and confusion matrix

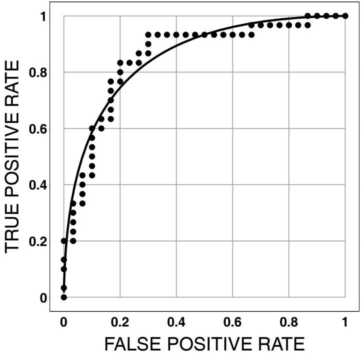
	Result of classification for fixed $\vartheta = \vartheta_0$	
Reality	**$\mathbf{g} \in \mathcal{C}_0$ (negatives)**	**$\mathbf{g} \in \mathcal{C}_1$ (positives)**
$\mathbf{g} \in \mathcal{C}_0$ (negatives)	true negatives $\mathrm{TN}(\vartheta_0)$	false positives $\mathrm{FP}(\vartheta_0)$
$\mathbf{g} \in \mathcal{C}_1$ (positives)	false negatives $\mathrm{FN}(\vartheta_0)$	true positives $\mathrm{TP}(\vartheta_0)$

False positive rate

$$\mathrm{FPR} \equiv \frac{\mathrm{FP}(\vartheta_0)}{\mathrm{TN}(\vartheta_0) + \mathrm{FP}(\vartheta_0)}$$

True positive rate

$$\mathrm{TPR} \equiv \frac{\mathrm{TP}(\vartheta_0)}{\mathrm{TP}(\vartheta_0) + \mathrm{FN}(\vartheta_0)}$$



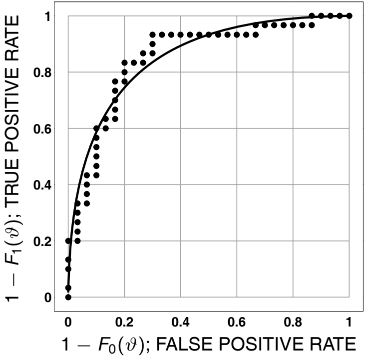
ROC curves and confusion matrix (cont.)

False positive rate

$$\mathrm{FPR} = \frac{\mathrm{FP}(\vartheta_0)}{\mathrm{TN}(\vartheta_0) + \mathrm{FP}(\vartheta_0)}$$

True positive rate

$$\mathrm{TPR} = \frac{\mathrm{TP}(\vartheta_0)}{\mathrm{TP}(\vartheta_0) + \mathrm{FN}(\vartheta_0)}$$



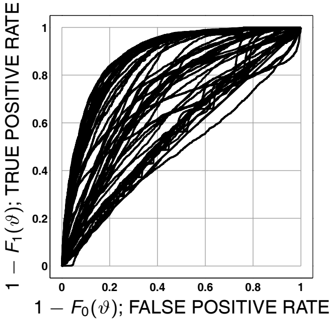
$$\frac{\mathrm{FP}(\vartheta)}{\mathrm{TN}(\vartheta) + \mathrm{FP}(\vartheta)} = \frac{1}{n_0}\mathrm{FP}(\vartheta) = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{I}(Y_{0i} > t) = 1 - \widehat{F}_0(\vartheta), \qquad \forall \vartheta \in \mathbb{R},$$

$$\frac{\mathrm{TP}(\vartheta)}{\mathrm{TP}(\vartheta) + \mathrm{FN}(\vartheta)} = \frac{1}{n_1}\mathrm{TP}(\vartheta) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}(Y_{1i} > t) = 1 - \widehat{F}_1(\vartheta), \qquad \forall \vartheta \in \mathbb{R}.$$

ÚFAL ROC curves

Linguistic collocation measures

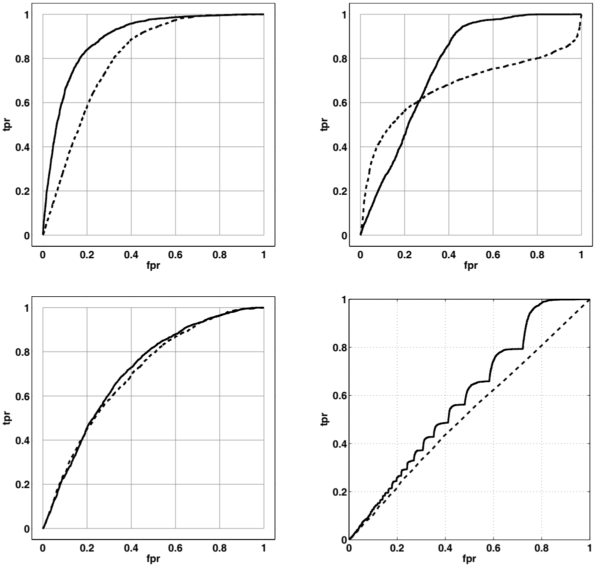
Linguist in ÚFAL use 86 different association measures for collocation extraction.



Tasks for statisticians

- How to compare these measures?
- How to detect groups (clusters) of collocation measures that behave analogously?

Typical linguistic ROC curves



Basic situation – reminder

Setup

- Diagnostic variable Y with conditional cdfs $F_0(\vartheta)$ and $F_1(\vartheta)$, i.e. $F_k(\vartheta) = \mathrm{P}(Y \le \vartheta \,|\, \mathcal{C}_k) \equiv \mathrm{P}(Y_k \le \vartheta),\; k = 0, 1$
- Y_0 and Y_1 follow continuous distributions with densities $f_0(\vartheta)$ and $f_1(\vartheta)$ such that $f_0(\vartheta) > 0$, $f_1(\vartheta) > 0$ on the same interval $\mathcal{I}_Y \subseteq \mathbb{R}$.
- Y_0 and Y_1 are independent.

Theoretical ROC curve is the range of

$$\begin{aligned} \varrho(\,\cdot\,; F_0, F_1) :\; \mathbb{R} &\to [0,1] \times [0,1] \\ \vartheta &\mapsto \big[1 - F_0(\vartheta), 1 - F_1(\vartheta)\big] \end{aligned}$$

$$\mathrm{ROC}_Y = \left\{ \boldsymbol{r} \in [0,1]^2 : \exists \vartheta \in \mathbb{R} \quad \varrho(\vartheta; F_0, F_1) = \boldsymbol{r} \right\}$$

Equivalence test – setting for two ROC curves

Two ROC curves

- $\mathrm{ROC}_Y = \left\{ \boldsymbol{r} \in [0,1]^2 : \exists \vartheta \in \mathbb{R} \quad \varrho(\vartheta; F_0, F_1) = \boldsymbol{r} \right\}$
- $\mathrm{ROC}_Z = \left\{ \boldsymbol{r} \in [0,1]^2 : \exists \vartheta \in \mathbb{R} \quad \varrho(\vartheta; G_0, G_1) = \boldsymbol{r} \right\}$

Observed data

- n_0 objects from $\mathcal{C}_0$ and n_1 objects from $\mathcal{C}_1$
- For each object two (different) measures Y and Z are evaluated
- It yields samples $Y_{01}, \ldots, Y_{0n_0}$ distributed according to $F_0(\vartheta)$, and $Y_{11}, \ldots, Y_{1n_1}$ distributed according to $F_1(\vartheta)$
- Analogously, $Z_{01}, \ldots, Z_{0n_0}$ – each follows $G_0(\vartheta)$, and $Z_{11}, \ldots, Z_{1n_1} \sim G_1(\vartheta)$

Equivalence test – idea

Equivalence of two ROC curves

for us means that for any $\boldsymbol{r}_Y \in \mathrm{ROC}_Y$ exists $\boldsymbol{r}_Z \in \mathrm{ROC}_Z$ such that

$$\boldsymbol{r}_Y = \boldsymbol{r}_Z$$

i.e.

$$\mathrm{ROC}_Y \equiv \mathrm{ROC}_Z \quad \Longleftrightarrow \quad \forall t_Y \in \mathcal{I}_Y \; \exists \, t_Z \in \mathcal{I}_Z : \; F_0(t_Y) = G_0(t_Z) \text{ and } F_1(t_Y) = G_1(t_Z)$$

Transformation function

Define $\tau_0, \tau_1 : \mathcal{I}_Y \to \mathcal{I}_Z$ such that

$$\tau_0(\vartheta) = G_0^{-1}\big(F_0(\vartheta)\big) \quad \text{and} \quad \tau_1(\vartheta) = G_1^{-1}\big(F_1(\vartheta)\big) \quad \forall \vartheta \in \mathcal{I}_Y.$$

Equivalence test – formal definition

Hypothesis

ROC curves are equivalent if and only if $\tau_0(\vartheta) \equiv \tau_1(\vartheta)$, i.e.

$$\mathsf{H} : \forall \vartheta \in \mathcal{I}_Y \qquad \tau_0(\vartheta) = \tau_1(\vartheta),$$

Alternative

We aim to test H against the alternative

$$\mathsf{A} : \exists\, \widetilde{\mathcal{I}}_Y \subseteq \mathcal{I}_Y, \widetilde{\mathcal{I}}_Y \neq \emptyset \quad \tau_0(\vartheta) \neq \tau_1(\vartheta) \quad \forall\, \vartheta \in \widetilde{\mathcal{I}}_Y$$

$$\tau_0(\vartheta) = G_0^{-1}\big(F_0(\vartheta)\big) \quad \text{and} \quad \tau_1(\vartheta) = G_1^{-1}\big(F_1(\vartheta)\big) \quad \forall \vartheta \in \mathcal{I}_Y.$$

Equivalence test – test statistic

Test statistic

$$T = n \int_{\mathcal{I}_Y^*} \left(\widehat{\tau}_0(\vartheta) - \widehat{\tau}_1(\vartheta)\right)^2 d\vartheta,$$

$$\widehat{\tau}_0(\vartheta) = \widehat{G}_0^{-1}\left(\widehat{F}_0(\vartheta)\right), \quad \widehat{\tau}_1(\vartheta) = \widehat{G}_1^{-1}\left(\widehat{F}_1(\vartheta)\right), \quad \forall \vartheta \in \mathcal{I}_Y,$$

$\widehat{F}_k(\vartheta)$ and $\widehat{G}_k(\vartheta)$, $k = 0, 1$, denote the empirical distribution functions,

$$\widehat{G}_k^{-1}(u) = \inf\left\{t : \widehat{G}_k(\vartheta) > u\right\}, \; k = 0, 1,$$

and closed interval $\mathcal{I}_Y^* \subseteq \mathcal{I}_Y$ is chosen such that

$$0 < g_0\left(\tau_0(\vartheta)\right) < \infty, \quad 0 < g_1\left(\tau_1(\vartheta)\right) < \infty, \quad \forall \vartheta \in \mathcal{I}_Y^*.$$

Equivalence test – test statistic

Theorem

Under the null hypothesis, the test statistic T converges weakly to the infinite weighted sums of independent χ^2_1 variables $\eta_1^2, \eta_2^2, \ldots$

$$T \xrightarrow[n\to\infty]{w} T^{\boldsymbol{B}} = \sum_{j=1}^{\infty} \lambda_j \eta_j^2,$$

where $\{\lambda_j\}$ represent the eigenvalues of the covariance operator of the zero mean gaussian process $B(t)$ with the covariance structure

$$\operatorname{cov}\big(B(s), B(t)\big) = c_0 \frac{F_0(s)\big(1 - F_0(t)\big)}{g_0\big(\tau_0(s)\big) g_0\big(\tau_0(t)\big)} + c_1 \frac{F_1(s)\big(1 - F_1(t)\big)}{g_1\big(\tau_1(s)\big) g_1\big(\tau_1(t)\big)}, \; s \le t,$$

c_0, c_1 are positive constants.

Equivalence test – critical values

To obtain (asymptotic) critical values we need

1. to estimate eigenvalues $\{\lambda_j\}$
2. to evaluate the distribution function of a weighted sum of χ_1^2 variables

To estimate the covariance structure and λ_j's

$$\widehat{\text{cov}}\left(B(s), B(t)\right) = c_0 \frac{\widehat{F}_0(s)\big(1-\widehat{F}_0(t)\big)}{\widetilde{g}_0\left(\widehat{\tau}_0(s)\right)\widetilde{g}_0\left(\widehat{\tau}_0(t)\right)} + c_1 \frac{\widehat{F}_1(s)\big(1-\widehat{F}_1(t)\big)}{\widetilde{g}_1\left(\widehat{\tau}_1(s)\right)\widetilde{g}_1\left(\widehat{\tau}_1(t)\right)},$$

for $s, t \in \{t_1, \ldots, t_p\} \subset \mathcal{I}_Y$, with $\widetilde{g}_k, k = 0, 1$, being density kernel estimators. The spectral decomposition of the matrix

$$\left(\widehat{\text{cov}}\left(B(t_i), B(t_j)\right)\right)_{i,j=1}^{p}$$

Equivalence test – Monte Carlo critical values

Trimming and Monte Carlo

Suppose that eigenvalues $\lambda_1, \ldots, \lambda_J$ are estimated. It allows an approximation of $T^{\boldsymbol{B}}$ by its first J estimated components

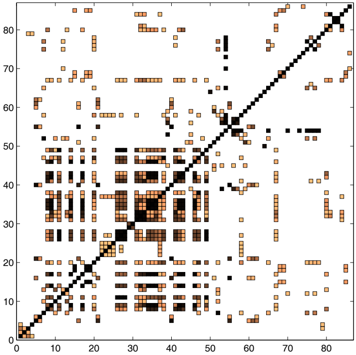
$$T^{\boldsymbol{B}} \approx \sum_{j=1}^{J} \widehat{\lambda}_j \eta_j^2 = S^J.$$

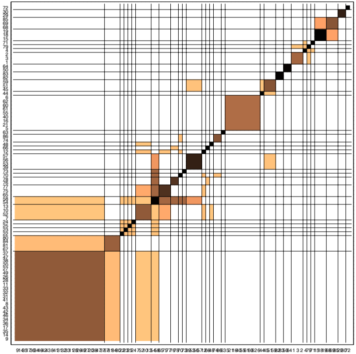
As distribution of S^J is not explicitly known, we perform Monte Carlo simulations in order to obtain the corresponding quantiles.

Proximity matrix for linguistic association measures

Application on linguistic association measures

We applied our test on all pairs of 86 collocation association measures and used $1 - p$-value as the proximity distance between two ROC curves.

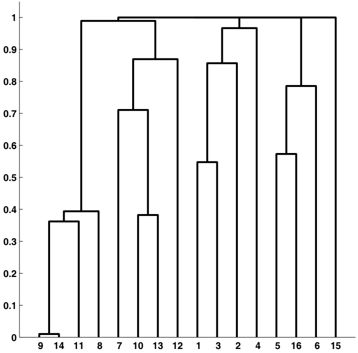


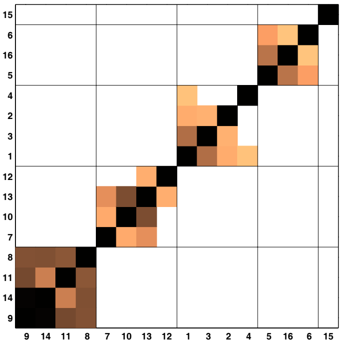


Collocation extraction – selected results

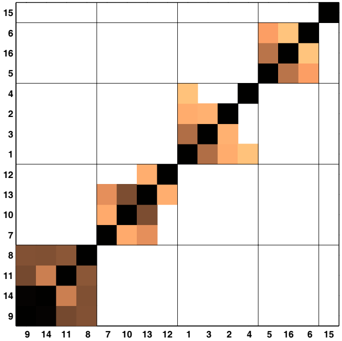
Proximity matrix and dendrogram

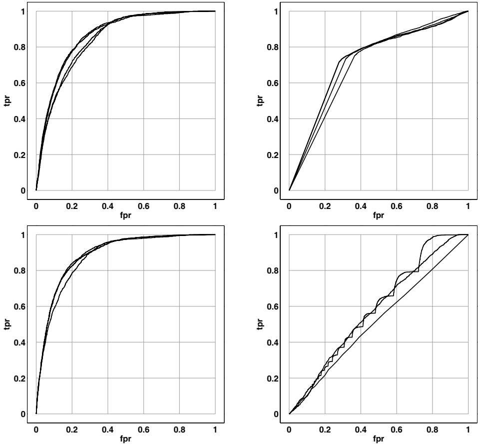
Rearranged (permuted) proximity matrix and the corresponding dendrogram, both providing insight ideas on natural similarity clusters of the observed ROC curves.



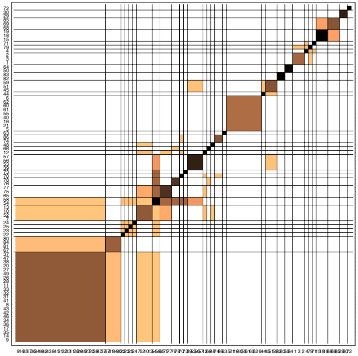


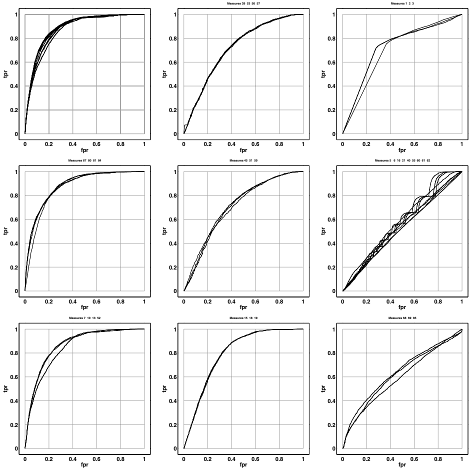
Equivalent classes of selected collocation rules





Equivalent classes of collocation rules





Lingustic problem – summary

- ROC curves are useful to display overall performance of a binary classifier
- ROC curve has a theoretical definition
- Statistical theory helps to understand properties of ROC curves and derive new analytical methods
- Our test is essentially based on a proper definition of a ROC curve. However, even straightforward ideas lead to quite complicated theoretical tasks
- Our test can be used to cluster ROC curves
- Clusters may serve to construct a superclassifer more efficient than individual measures

DĚKUJI