Inductive Constructed a Set of PROSPECTOR-like Rules for Classification of Literary Works

Sergey P. Chistjakov
Karelian Research Centre of Russian Academy of Science, Russia,
chistiakov@krc.karelia.ru

Tatyana G. Surovtsova
Petrozavodsk State University, Russia,
tsurovceva@psu.karelia.ru

Abstract

The method based on inductive constructed a set of PROSPECTOR-like rules we used for literary works classification.

Description of syntactic and morphological features of texts according to grammar of Russian calculated by means of expert system, an integrated component of information retrieval system ”SMALT”.

Inductive generation rules and recognition literary works completed by means of programs for inductive construction of the knowledge base “STATCOP”.

Prospector–like knowledge bases and classifiers

We begin with some notation. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be some nominal attributes vector and $\mathrm{X} = \mathrm{X}_1 \times \mathrm{X}_2 \times \ldots \times \mathrm{X}_n$, where $\mathrm{X}_i = \{x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$, $i = 1, 2, \ldots, n$ is a set of possible values of the attribute X_i. Also let Y be a class attribute with a set of possible values $D = \{0, 1, \ldots, k-1\}$. Suppose that the unknown joint distribution $P(\mathbf{x}, y)$ of the attributes $X_1, X_2, \ldots, X_n, Y$ exists. Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ be a training set.

We consider a sets of the rules of the form "IF $\langle$ *premise* $\rangle$ THEN $\langle$ *conclusion* $\rangle$ $\langle$ *weight* $w\rangle$" , where premise C (*combination*) has the form

$$C = \{X_{\alpha_1} = x_{\alpha_1\beta_1}\} \cap \{X_{\alpha_2} = x_{\alpha_2\beta_2}\} \cap \ldots \cap \{X_{\alpha_r} = x_{\alpha_r\beta_r}\}$$

and conclusion $C_i^* = \{Y = i\}$, $i \in D$. Weight $w \in (0, 1)$ is a quantitative measure reflecting the influence of the premise on the conclusion. The value r is called the length of combination C and denote by $length(C)$. Also denote by $\|C\|_{\mathcal{D}}$ the number of examples of the training set such that premise C is true. The rules of such form we shall denote $C \Rightarrow C_i^*$ $\langle w \rangle$. Let $C_1 \Rightarrow C_i^*$ $\langle w_1 \rangle$ and $C_2 \Rightarrow C_i^*$ $\langle w_2 \rangle$ be some rules with the same conclusion C_i^*. Then following weight combining function may be defined [3]:

$$w_1 \oplus w_2 = \frac{w_1 w_2}{w_1 w_2 + (1 - w_1)(1 - w_2)}.$$

Let $\mathcal{R}$ be some set of rules. Then for any $i \in D$ and combination C the following *composed weight* may be evaluate: $W(C_i^* | C, \mathcal{R}) = \bigoplus_{\alpha} w_\alpha$, where weight combining function $\oplus$ is applied to the weights w_α of all rules $C' \Rightarrow C_i^* \ \langle w_\alpha \rangle$ containing in $\mathcal{R}$ such that the premise C' follows from the premise C. Note that the composed weight $W(C_i^* | C, \mathcal{R})$ actually is an estimate of conditional probability $\mathbf{P}(C_i^* | C)$. Then the set of rules $\mathcal{R}$ induce some classifier $f_{\mathcal{R}} : \mathrm{X} \to D$ such that for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathrm{X}$

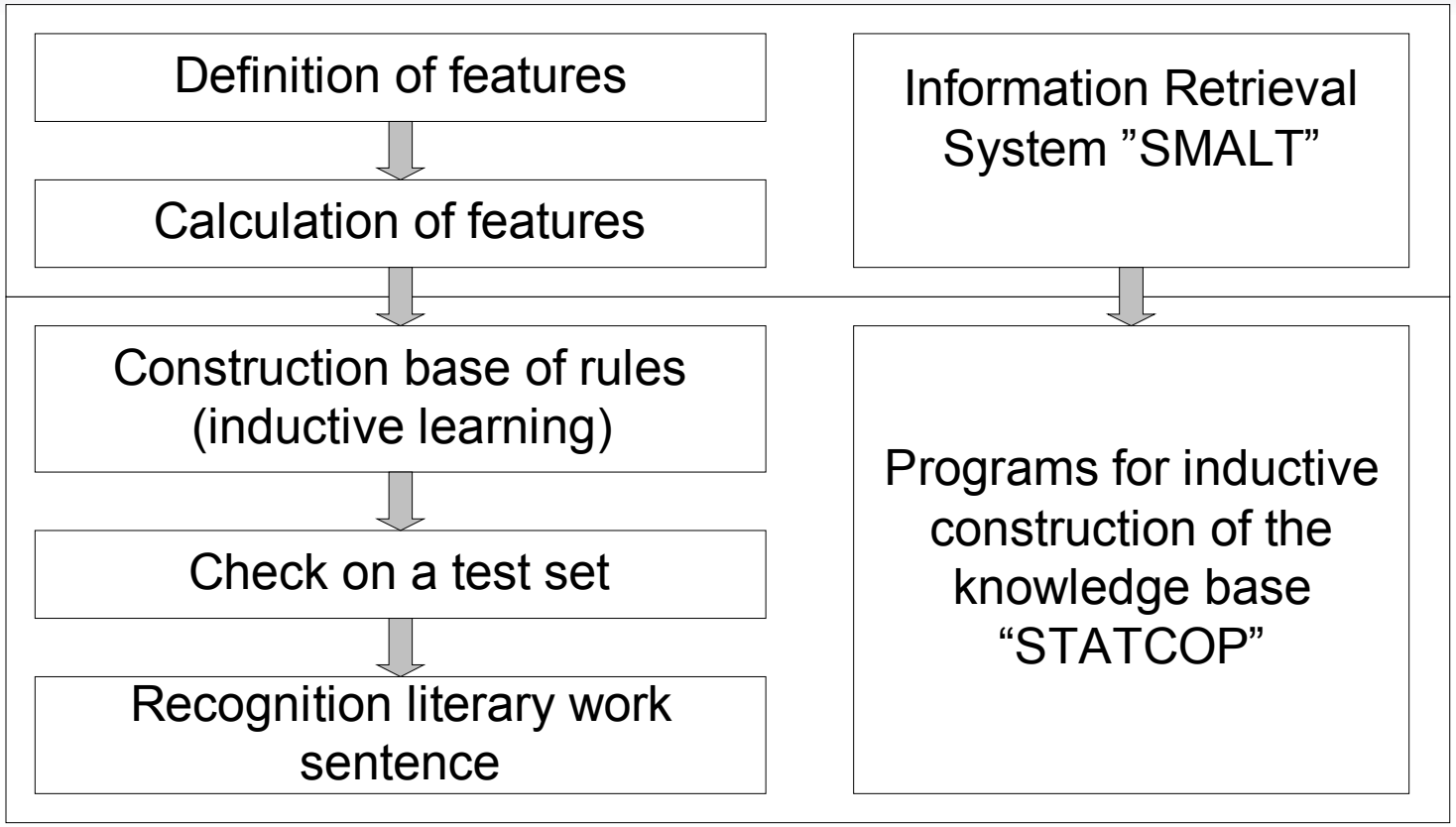
$$f_{\mathcal{R}}(\mathbf{x}) = \arg\max_i W(C_i^* | C(\mathbf{x}), \mathcal{R}),$$

where $C(\mathbf{x}) = \{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}$. Also, let $\mathbf{K}$ be some statistical test for testing of the null hypothesis $H_0 : \mathbf{P}(C_i^* | C) = \theta_0$ against the two-side alternative $H_1 : \mathbf{P}(C_i^* | C) \neq \theta_0$ and Δ is some set of the *admissible* implications $\{C \Rightarrow C_i^*\}$. For example, Δ may includes all implications such that $length(C) \leq 3$ and $\|C\|_{\mathcal{D}} > F_{min}$.

Definition 1. [1]*Let training set $\mathcal{D}$, weight combining function $\oplus$, statistical test $\mathbf{K}$, some set Δ of admissible implications $C \Rightarrow C_i^*$, and class attribute Y are given. The set of rules $\mathcal{KB}(\Delta) = \mathcal{KB}(\mathcal{D}, \oplus, \mathbf{K}, \Delta, Y)$ is called knowledge base (with respect to $\mathcal{D}, \oplus, \mathbf{K}, \Delta, Y$) if for any premise C such that implication $C \Rightarrow C^*$ is admissible statistical test $\mathbf{K}$ does not reject hypothesis $H_0 : \mathbf{P}(C^* | C) = W(C^* | C, \mathcal{KB})$ against two–side alternative. Knowledge base $\mathcal{KB}(\Delta)$ is called minimal if $\mathcal{KB}$ is subset of any set of rules $\mathcal{R}(\Delta)$ which is a knowledge base with respect to $\mathcal{D}, \oplus, \mathbf{K}, \Delta, Y$.*

By definition 1, it follows that the knowledge base in consideration in fact is the probably–statistical model of the condition probabilities family $\{\mathbf{P}(C_i^* | C),\ i \in D\}$ such that implications $C \Rightarrow C_i^*$ are admissible.

Description of a Experiment



Training Set and Features Vector

Training Set

Code	Author	Title	Quantity of sentences
D1	Ф.М. Достоевский (F.M. Dostoevsky)	"Ряд статей о русской литературе. Введение."	669
A1	М.М. Достоевский (M.M. Dostoevsky)	“Жуковский и романтизм."	225

Features Vector, $X=\{X_1, X_2, \ldots X_{20}\}$

- Parts of speech on various positions of the sentence
- Relative quantity of various parts of speech of the sentence
- Syntactic characteristics of the sentence
- …

Result of Experiment

Rules (99 in all)

IF (relative quantity of particles in the sentence > 14.945) THEN (D1) {weight = 0.855}

IF (average length of a word in letters in the sentence <= 4.845) THEN (D1) {weight = 0.818}

IF (relative quantity of nouns in the sentence <= 12.31) THEN (D1) {weight = 0.739}

IF (25.87 < relative quantity of nouns in the sentence <= 33.94) THEN (D1) {weight = 0.355}

IF (part of speech in a penultimate position on the sentence = adjective) THEN (D1) {weight = 0.326}

IF (average length of a word in letters in the sentence > 6.41) THEN (D1) {weight = 0.267}

Result of sentences classification

	A1	D1
A1	130	95
D1	26	199

Empirical Risk = 0,269

Conclusions

Inductive constructed a set of PROSPECTOR-like rules can be used for classification of literary works.

We continue researches and check on more extensive test set.

References

1 Berka P., Ivanek J. Automated Knowledge Acquisition for PROSPECTOR-like Expert Systems. Proceeding ECML'94, Springer, 1994. - pp. 339-342.

2 Chistjakov S.P. Ob avtomatizacii postroeniya baz znanii expertnuh sistem // Obozrenie prikladnoi i promushlennoi matematiki, t. 8, vyp. 1, 2001. – pp. 375-376.

3 Hajek P. (1985). Combining Functions for Certainty Factors in Consulting Systems. Int. J. Man-Machine Studies, 1985. - Vol. 22, pp. 59-76.

4 Juola P., Sofko J., Brennan P. *A Prototype for Authorship Attribution Studies* // Literary and Linguistic Computing, Vol. 21, No. 2, 2006, pp. 69-178.

4 Rogov A.A., Sidorov Yu. Vl. *Statistical and Information-calculating Support of the Authorship Attribution of the Literary Works*. Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods: Proc. of the Sixth International Conference (September 10-14, 2001, Minsk). Vol.2: K-S/ Edited by Prof. Dr. S. Aivazian, Prof. Dr. Yu. Kharin and Prof. Dr. H. Rieder. Minsk: BSU, 2001. – P. 187-192.