as a manuscript

Masyutin Alexey

INTERVAL PATTERN STRUCTURES RANDOMIZED ALGORITHMS FOR CLASSIFICATION AND REGRESSION TASKS IN CREDIT RISK MANAGEMENT

PhD Dissertation Summary

for the purpose of obtaining
Philosophy Doctor in Computer Science HSE

Moscow — 2018

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor:	Sergei O. Kuznetsov, Doctor of Science, professor of
:	National Research University Higher School of Economics

PhD Dissertation Relevance. The development of information technologies creates a much tougher competitive environment for banks and credit institutions. For example, since January 2017, Russia's two largest telecommunications companies have begun to provide loans to their customers, although they have never developed this business line before[1]. The provision of such services on the part of non-financial companies was made possible precisely through the introduction of a modern IT infrastructure to store a large amount of data on customers and use mathematical modeling techniques to assess the creditworthiness of customers. The international leader in the sphere of strategic consulting services, McKinsey, publishes studies according to which by 2025 the world banking industry will undergo significant changes. One of the main reasons for the transformation is the digitalization of banking services, the rapidly growing volume of data on customers and their operations, the emergence of new types of risk associated with the use of machine learning in decision-making.

Mathematical modeling in the banking sector finds one of its main applications in the field of risk management. A prerequisite for effective risk management is the ability to adequately assess the probability and magnitude of the risk. At the moment, risk assessment problems are approached by a wide range of statistical tools, such as scorecards, rating models, loss given default modeling. At the same time one of key reasons for the numerous reviews of licenses is, among other things, the low quality of the loan portfolio; inadequate estimates of the probability of default and / or loss given default[2] (for example, Probusinessbank, Tatfondbank in Russia). So, in the press releases of the Central Bank of Russia, one can find following statements: "with unsatisfactory asset quality credit institution inadequately assessed the risks", "as a result of expected loss calculation given the risks accepted, the credit institution completely lost its capital", etc.

With the increase in the volume of customer data, standard methods of classification and regression are yielding lower accuracy in comparison to more complex algorithms, such as random forests, support vector machines, and decision trees. Nevertheless, risk management requires the property of interpretability of the predictions received by the model, which in the case of complex algorithms is usually impossible. In addition, there are certain prescriptions of the Bank of Russia that regulate the process of assessing credit risk on the basis of mathematical models, and banks undergo a detailed procedure for validating the use of models for risk assessment, a significant part of which is to test the stability of the model and its business logic.

---

[1] http://www.rbc.ru/finances/11/01/2017/587500529a794767fa723fa8, above-mentioned companies are Beeline and MTS.

[2] http://www.cbr.ru/press/PR/?file=12082015_085127ik2015-08-12t08_46_23.htm

This PhD Dissertation solves the problem for interpreted decision rules that outperform classification and regression methods conventional in the banking sphere, such as scorecards and decision trees, but at the same time preserve the interpretability of the forecasts produced.

This is achieved by formal concept analysis (FCA) and interval pattern structures technique. Several new definitions and modifications were introduced so that one could be able to perform continuous target variable prediction via interval pattern structures and to handle datasets with considerable number of observations.

Object of Research is interpretable mathematical models for probability of default and loss given default estimation.

PhD Dissertation Goal is the development of innovative methods for credit scoring and loss given default estimation which provide higher accuracy in comparison to credit scorecards and decision trees, provided the methods are interpretable from business logic perspective.

The two main **novelties** of PhD Dissertation are as follows:

1. Query-Based Classification Algorithm which provides randomized procedure to solve credit scoring problem for datasets with large number of observations;
2. Adopting tools of formal concept analysis to regression task (i.e. case of continuous target variable) by developing Query Based Regression Algorithm. The algorithm allows one to solve loss given default problem.

PhD Dissertation offers data analysis algorithms that have accuracy superior to simple algorithms widely adopted within the banks (such as logistic regression, decision trees and scorecards) and still maintain the property of interpretability in sense that they provide a decision maker with a set of rules applicable to the borrower creditworthiness assessment. In order to achieve this goal several novelties within the methods of formal concept analysis (FCA) and interval pattern structures were introduced. The reasons why FCA methods are suitable for credit risk assessment under the interpretability requirements are discussed in the main text of the dissertation. The novelty brought to the well-developed tools of FCA consists of two parts.

The first one is that FCA is adopted to credit scoring problem based on numerical data with the step of concept lattice construction being omitted (query-based classification or "lazy" classification). This allows one to work with the datasets with large number of observations which is vital for banks as soon as historical data is typically large.

The second one is that we introduce a modification to FCA method based on interval pattern structures which allows one to solve regression problem. To our knowledge FCA methods were not applicable to such type of data analysis problem

before. The crucial difference in regression problem is that the target variable is distributed continuously. The proposed method is designed to solve loss given default prediction problem.

The **Practical Value** is justified by experiments on the comparative evaluation of the different algorithms for classification and regression problems on real bank data as well as open data. All proposed methods are implemented as software packages designed to solve credit scoring problems.

Proposed methods and algorithms were applied as a test project to the company's two client datasets of top-10 Russian bank and results of accuracy calculations and benchmark analysis are provided in the dissertation.

The reliability of the obtained results is justified by application of mathematical models and methods, by experiments comparing the results of applying the traditional methods developed.

Publications.

First-tier publications:

1. Masyutin A., Kashnitsky Y. Query-Based Versus Tree-Based Classification: Application to Banking Data // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, 10352 LNAI, pp. 664-673.

Second-tier publications:

2. Masyutin A. Credit scoring based on social network data // Business Informatics, No. 3 (33), 2015, pp. 15–23.
3. Masyutin A. Alternative Ways for Loss-Given-Default Estimation in Retail Banking //Analysis of Images, Social Networks and Texts, 2014, Volume 436 of the series Communications in Computer and Information Science, pp. 152-162.
4. Masyutin A., Kashnitsky Y., Sergei O. Kuznetsov. Lazy Classication with Interval Pattern Structures: Application to Credit Scoring, in: Proceedings of the International Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI at IJCAI 2015), CEUR Workshop proceedings, Vol.1430, p. 43-54.
5. Masyutin A., Sergei O. Kuznetsov. Continuous Target Variable Prediction with Augmented Interval Pattern Structures: A Lazy Algorithm // Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, pp.273-284, 2016.

Other publications:

1. Masyutin A., Borisyuk V. Fraud Detection Scorecard // Risk-management in credit organization, No. 2/2015. (*in Russian*)

Discussions at conferences and seminars.

1. 3th International Conference - Analysis of Images, Social Networks and Texts, AIST'2014, Ekaterinburg, Russia.

Subject: "Alternative Ways for Loss-Given-Default Estimation in Retail Banking";

2. 24th International Conference on Artificial Intelligence, Workshop "What can FCA do for Artificial Intelligence?", FCA4AI at IJCAI 2015, Buenos Aires, Argentina.

Subject: "Lazy Classification with Interval Pattern Structures: Application to Credit Scoring;

3. Research Seminar of the Graduate School of Computer Science, CS HSE, November 12, 2015.

Subject: "Integrated risk management and machine learning tasks", (https://aspirantura.hse.ru/cs/announcements/164947571.html);

4. 13th International Conference on the Concept of Lattices and Their Applications, CLA'2016. Moscow, Russia

Subject: "Continuous Target Variable Prediction with Augmented Interval Pattern Structures: A Lazy Algorithm";

5. Database Technologies, 2016, Moscow, Russia (https://www.osp.ru/iz/tbd_dbms)

Subject: "Credit scoring based on the analysis of formal concepts: personal models";

6. 23rd International Symposium on Methodologies for Intelligent Systems, ISMIS'2017, Warsaw, Poland.

Subject: "Query-based versus tree-based classification: application to banking data";

7. Seminar of ISSA, CS HSE, September 28, 2017.

Subject: "Classification on demand based on interval descriptions in credit risk management tasks" (https://cs.hse.ru/ai/issa/announcements/209647835.html).

PhD Dissertation Contents. This dissertation consists of 4 sections, conclusion and bibliography of 89 resources.

The **first section** (introduction) covers the dissertation relevance, the problems and tasks of the research, the subject of the research. Also, objectives of the work are determined, the main results are stated, theoretical and practical significance of the work is discussed.

The **second section** describes the history of mathematical modeling within banking industry. It covers the key role modeling is playing in risk management and reviews widespread statistical algorithms used for classification and regression tasks. In context of credit risk assessment two parameters are emphasized: probability of default (PD), loss given default (LGD). From data science standpoint PD estimation is a binary classification problem, and LGD estimation is regression problem. Trade-off between accuracy of prediction versus model interpretability is emphasized as soon as some regulators require banks to be able to provide reject reasons for borrowers and also when central banks examine the bank models they are willing to understand economic intuition behind them to prove the models are going to show expected and stable performance.

Loan default prediction with the use of scorecards is discussed as soon as this method is widely adopted within banking industry and is used as benchmark for "white-box" models thereafter. The raw factors weight-of-evidence (WOE) transformation is designed to account for outliers and non-monotonous dependencies in an adequate way before feeding data into logistic classifier.

"Black-box" models are discussed with an example of neural networks which are opposed to transparent models which provide user with understanding why the algorithm predicts particular probabilities of default for client.

The **third section** contains the first novelty: application of formal concept analysis (FCA) to classification problem with datasets of large number of observations. Basic FCA definitions are provided (*pattern structure, meet operator, derivation operator, pattern intent and extent*). New definitions for α-weak premises[3] are provided.

Suppose we have a set of *positive examples* G_+ (objects of positive class) and a set of *negative examples* G_- (objects of negative class), $G_+ \cap G_- = \emptyset$, и $G_+ \cup G_- = G$. Let the description set is denoted by D, which consists of tuples with intervals as its elements, i.e. $D = \{([a_1; b_1], \dots, [a_{K,}; b_{K,}]) \mid \forall i: a_i, b_i \in R\}$, where K is dimensionality of attribute space. For example, for K=3 one can provide the following element of D: d = ([1;2], [-0.5;0.3], [150;340]).

[3] also known as classifiers, hypotheses

Let us provide mapping $\delta: G \to D$ such that for $g \in G$: $\delta(g) = ([a_1; a_1], \dots, [a_K; a_K])$, i.e. each object has its own description as a point in *K*-dimensional real number space.

For two descriptions $d_1, d_2 \in D$, $d_1 = ([a_1; b_1], \dots, [a_K; b_K])$ and $d_2 = ([m_1; n_1], \dots, [m_K; n_K])$ *meet operation* $\sqcap$ is defined:

$$d_1 \sqcap d_2 = ([\min(a_1, m_1); \max(b_1, n_1)], \dots, [\min(a_K, m_K); \max(b_K, n_K)])$$

If $d_1 \sqcap d_2 = d_1$, then it is denoted that $d_1 \sqsubseteq d_2$

Interval pattern structure is a triplet $(G, \underline{D}, \delta)$, где $\underline{D} = (D, \sqcap)$, i.e. a set of objects with a set of possible descriptions, meet operation $\sqcap$ and a mapping δ.

Also we define a mapping from set of objects *G* to description set *D* and vice versa, denoting it with $^\diamond$:

$$A^\diamond = \sqcap_{g \in A} \delta(g) \text{ for } A \subseteq G,$$
$$d^\diamond = \{g \in G \mid d \sqsubseteq \delta(g)\} \text{ for } d \in D.$$

New definitions of α-weak premises are introduced. Description $d_+ \in D$ is called an *α-weak positive premise* if:

$$\frac{|d_+^\diamond \cap G_-|}{|G_-|} \leq \alpha, \text{ and } \exists A \subseteq G_+ \colon\ d_+ \sqsubseteq A^\diamond$$

Description $d_- \in D$ is called an *α-weak negative premise* if:

$$\frac{|d_-^\diamond \cap G_+|}{|G_+|} \leq \alpha, \text{ and } \exists A \subseteq G_- \colon\ d_- \sqsubseteq A^\diamond$$

Query-based classification algorithm ("lazy classification") is introduced. The algorithm takes set of positive and negative examples ($G_+$and G_-), set of test objects G_{test} with corresponding descriptions and mapping δ as input. The output of the algorithm is a real number $\Delta \in R$ assigned for each test object from $g_{test} \in G_{test}$. This number Δ serves as a credit score and allows one to build cutoff decision rules such as "if $\Delta > x$ then g_{test} belongs to positive class". The idea behind the algorithm is to check whether it is the set of positive or negative examples the test object is more similar to. The similarity is defined as a *total support* of α-weak positive (negative) premises that contain the description of test object. The *support* of an α-weak positive description of d_+ is called $|d_+^\diamond \cap G_+|$, that is, the number of objects from the set of positive examples G_+ satisfying the description of d_+. The support of an α-weak negative description of d_- is called $|d_-^\diamond \cap G_-|$, that is, number of objects from the set of negative examples G_- satisfying the description of d_-. Let there be *p* α-weak positive descriptions and *n* α-weak negative descriptions, all of them contain the description of the test object $\delta(g_{test})$, i.e. $\forall i = 1, \dots, p \colon d_{+i} \sqsubseteq \delta(g_{test})$ and $\forall j = 1, \dots, n \colon d_{-j} \sqsubseteq \delta(g_{test})$. The *total support* for α-weak *positive* descriptions is $P = \sum_{i=1}^{p} |d_{+i}^\diamond \cap G_+|$, and the *total support* for α-weak *negative* descriptions is

$N = \sum_{j=1}^{n} |d_{-j}^{\circ} \cap G_{-}|$. Based on the value $\Delta = P - N$, an estimation is made whether the test object is more similar to objects from a set of positive or negative examples; it serves as credit score for the borrower's creditability assessment. The paper also considers other similarity measures and voting schemes based on α-weak descriptions (see section 3.4 of the dissertation).

The algorithm is an iterative procedure and uses three hyperparameters: *subsample size*, *number of iterations* and *alpha-threshold*. The first hyperparameter is a percentage of objects in a set of positive (negative) examples which are randomly extracted within each iteration. At each iteration the subsample is extracted from G_{-} and G_{+} and objects descriptions in subsample are intersected (⊓) with the description of test object g_{test}:

$$d = \delta(g_1) \sqcap \ldots \sqcap \delta(g_k) \sqcap \delta(g_{test})$$

where $k/|G| = subsample\_size$.

The number of times (*number of iterations*) we randomly extract a subsample from the set of examples is the second hyperparameter of the algorithm, which is also tuned through grid search. If d is not α-weak premise then it is ignored, if d is α-weak premise then d is saved in order to be used in classification of the test object later.

These steps are performed for each test object for positive and negative set of examples separately, producing a set of positive and negative α-weak premises. *The final output* of the algorithm is a difference between the total support for α-weak positive premises and the total support for α-weak negative premises for the test object. Based on this output we calculate model quality metrics that is widely used in credit scoring – Gini coefficient.

The algorithm is tested on both internal top-10 bank data and open Kaggle data. The positive set of examples is a set of loans where the target attribute is present. The target attribute in credit scoring is defined as more than 90 days of delinquency within the first 12 months after the loan origination. Each set of examples consists of 1000 objects in order that voting scheme concerned in the second section was applicable. The test dataset consists of 300 objects and is extracted from the same population as the sets of positive and negative examples. Attributes represent various metrics such as loan amount, term, rate, payment-to-income ratio, age of the borrower, undocumented-to-documented income, credit history metrics etc. The set of attributes used for the lazy classification trials contained 28 numerical attributes. In order to evaluate the accuracy of the classification Gini coefficient is calculated for every combination of hyperparameters based on 300 predictions on the test set. Gini coefficient is calculated based on the margin between the number of objects within positive

premises and negative ones. The margin is considered as an measure similar to score value in credit scorecards. Hyperparameter grid search is performed:

Gini coefficients for the hyperparameters grid search (QBCA)

		Subsample size								
Alpha-threshold	Number of iterations	0.1%	0.2%	0.3%	0.4%	0.5%	0.6%	0.7%	0.8%	0.9%
0.0%	100	40%	44%	39%	18%	1%	0%	0%	0%	0%
	150	35%	46%	35%	5%	0%	0%	0%	0%	0%
	200	42%	37%	36%	12%	5%	1%	0%	0%	0%
	500	39%	44%	44%	25%	6%	1%	0%	0%	0%
	1000	44%	47%	44%	41%	11%	3%	0%	0%	0%
	2000	44%	48%	46%	36%	17%	4%	0%	0%	0%
0.1%	100	33%	37%	40%	40%	44%	43%	34%	32%	34%
	150	41%	34%	33%	43%	41%	47%	41%	37%	37%
	200	40%	40%	34%	42%	51%	43%	44%	41%	36%
	500	37%	42%	47%	49%	51%	49%	43%	41%	34%
	1000	37%	42%	46%	48%	49%	48%	43%	43%	37%
	2000	39%	43%	45%	49%	51%	49%	46%	41%	38%
	5000	43%	40%	44%	49%	46%	50%	48%	38%	36%
0.2%	100	29%	38%	42%	32%	43%	37%	46%	43%	37%
	150	27%	42%	41%	41%	36%	47%	48%	45%	41%
	200	32%	40%	43%	42%	42%	49%	46%	47%	48%
	500	39%	46%	46%	48%	47%	48%	51%	48%	51%
	1000	41%	50%	48%	47%	49%	53%	52%	52%	47%
	2000	38%	48%	50%	48%	47%	53%	52%	53%	50%
0.3%	100	35%	38%	39%	42%	39%	45%	34%	45%	39%
	150	27%	43%	44%	42%	42%	39%	37%	40%	46%
	200	34%	46%	47%	45%	49%	47%	45%	45%	52%
	500	31%	45%	49%	50%	49%	46%	50%	51%	47%
	1000	37%	48%	49%	49%	49%	47%	52%	51%	51%
	2000	38%	46%	48%	51%	51%	50%	50%	52%	52%
	5000	40%	47%	46%	51%	52%	51%	49%	51%	53%
	10000	40%	44%	43%	46%	46%	48%	50%	52%	54%
	20000	40%	43%	42%	46%	47%	49%	50%	52%	53%

Gini coefficients for the hyperparameters grid search on specified area

		Subsample size					
Alpha-thresh-old	Number of iterations	1.0%	1.1%	1.2%	1.3%	1.4%	1.5%
0.3%	500	51%	49%	48%	43%	41%	38%
	1000	52%	51%	48%	45%	43%	39%
	2000	54%	53%	49%	47%	46%	38%
	5000	55%	52%	50%	47%	46%	40%
	10000	56%	53%	50%	47%	47%	40%
	20000	55%	53%	51%	46%	48%	41%

Query-based classification algorithm versus classical models adopted in banks and other benchmarks for top-10 bank data

	Gini on test sample
Logistic regression	47.38%
Scorecard (Logistic based on WOE-transformation)	51.89%
CART (minsize= 50)	54.75%
MLCA (s = 1%, a=0.3%, n=10000)	56.30%
AdaBoostClassifier	54.72%
KNeighborsClassifier	44.00%
NaiveBayes	48.91%
RandomForestClassifier	53.42%

As far as open data is concerned the algorithm was tested on Kaggle data of "Give Me Some Credit" contest held in 2012[4]. The data has a binary target variable (class label) whether the borrower defaulted or not.

Query-based classification algorithm versus benchmarks for Kaggle credit scoring open dataset

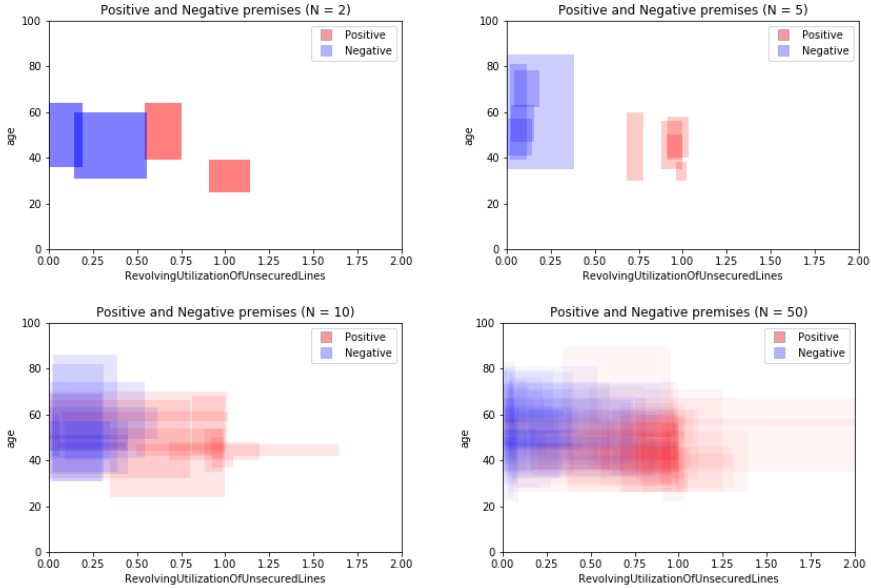
metric \algo	**Scorecard**	**QBCA**	**Xgboost**
Valid. Gini	0.5806	0.6624	0.708

[4] https://www.kaggle.com/c/GiveMeSomeCredit

Apart from accuracy measures sensitivity analysis is performed and algorithms properties are analyzed.

Also, visualization of collections of α-weak premises are presented that allows one to interpret the model outcome for the client. In effect, when considering a test object target class label (good or bad) the algorithms builds portraits of good and bad clients on historical data in multi-dimensional feature space.

Below are several examples of two-way areas are provided for different levels of number of iterations:



Positive premises are depicted in red and negative are in blue. To construct each positive premise, two objects from the set of positive examples were randomly extracted. Then the meet-operator was applied and a set of intervals was obtained. After that, only the intervals for two features were left. The same algorithm was performed for negative premises.

It is argued that this set of areas in feature space gives the understanding of why a particular borrower was considered of high or low credit risk by the model.

The **fourth section** contains the second novelty: adoption of FCA to regression problem (i.e. the target variable is distributed continuously). In order to

make FCA techniques applicable to this case new definitions of *augmented interval pattern structure* is given.

An *augmented interval pattern structure* is a quadruplet: $(G, \underline{D}, \delta, h)$, where G is a set of objects, D is a set of possible object descriptions, $d \in D$, and $\sqcap$ is a meet operator. Description d in credit scoring domain is a tuple which consist of two elements d_x and d_y (d_y – is an interval for *target attribute* $y \in R$, and d_x – a tuple of intervals for *explanatory attributes x,* which are supposed to predict target attribute *y*).

Let there be a mapping δ*:* $G \to D$ and additionally empirical distribution function $h \in H$, where H is a density functions family for target attribute. We will also use notation δ_x *u* δ_y, to distinguish between descriptions containing explanatory attributes and target attribute correspondingly. The meet operator $\sqcap$ definition is left unchanged.

Suppose, we have an arbitrary set of objects $A_0 \in G$, i.e.:

$$A_0 = \{g_1, g_2, \dots, g_J\},$$
$$\delta(g_j) = (\delta_x, \delta_y) = ([x_{1j}; x_{1j}], \dots, [x_{Kj}; x_{Kj}], [y_j; y_j])$$
$$\text{for } j = 1, \dots, J,$$

where J is a number of explanatory attributes. Then we define derivation operator $^\circ$ the following way:

$$A_0^\circ = (d_0, h_0)$$

where $d_0 = \{d_{x0}, d_{y0}\}$ and $d_{x0} = \delta_x(g_1) \sqcap \dots \sqcap \delta_x(g_J)$, and target attribute description is $d_{y0} = \delta_y(g_1) \sqcap \dots \sqcap \delta_y(g_J)$, which is in fact a single interval [y_{min}, y_{max}], and h_0 is mapping $d_{y0} \to [0; 1]$, i.e. empirical density distribution function of target attribute values in A_0:

$$h\big([\tau_{i-1}, \tau_i)\big) = \frac{\sum_{g \in A} 1_{[\tau_{i-1}, \tau_i) \sqsubseteq \delta_y(g)}}{|A|}, \forall i = 1, \dots, K$$

where $\tau_0 = y_{min}, \tau_K = y_{max}$, and $\Delta\tau_i = \tau_i - \tau_{i-1} = \frac{y_{max} - y_{min}}{K}$, 1 is indicator function.

We will use the composition of derivation operator $^\circ$ in a similar way, it was used with interval pattern structures, however it will return the image for description d_{0x} whatever target description d_{0y} and density function h are:

$$A_0^{\circ\circ} = (d_0, h_0)^\circ \stackrel{\text{def}}{=} d_{x0}^\circ = A_1$$

In order to approach target attribute prediction problem it will be useful to define *α-weak premise with ω-allowed dropout.* Augmented interval pattern structure $d = (d_x, d_y) \in D$ is called an α-weak premise with ω-allowed dropout, iff:

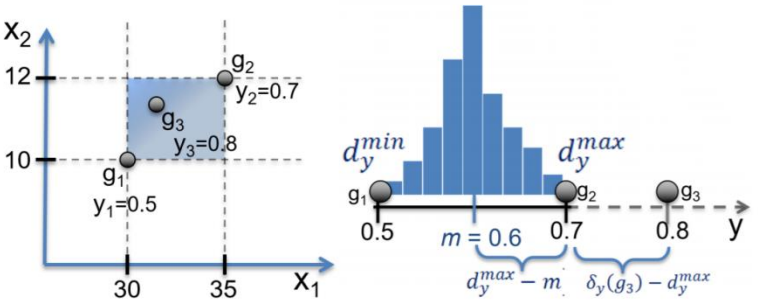
$$1 - \frac{|\{g \in A \mid d_y^{min} - \omega(m - d_y^{min}) \le \delta_y(g) \le d_y^{max} + \omega(d_y^{max} - m)\}|}{|A|} \le \alpha$$

where $\delta_y(g)$ is a value of target attribute for object g, and $A = d_x^{\circ}$, d_y is an interval $[d_y^{min}; d_y^{max}]$ for target attribute, and m is a median of empirical density distribution function h that describes target attribute values within interval d_y for objects from A.

Below we provide an example to understand how new definitions work.

Let object set be $G = \{g_1, g_2, g_3\}$ and description space consists of two explanatory attributes x_1, x_2 and one target attribute y:

Objects\Attributes	x_1	x_2	y
g_1	30	10	0.5
g_2	35	12	0.7
g_3	31.5	11.5	0.8



Let $A_0 = \{g_1, g_2\}$.

Then $\delta_x(g_1) = ([30; 30], [10; 10])$, $\delta_y(g_1) = [0.5; 0.5]$

$\delta_x(g_2) = ([35; 35], [12; 12])$, $\delta_y(g_2) = [0.7; 0.7]$

$d_0 = (d_{0x}, d_{0y})$

$d_{0x} = \delta_x(g_1) \sqcap \delta_x(g_2) = ([30; 35], [10; 12])$

$d_{0y} = \delta_y(g_1) \sqcap \delta_y(g_2) = [0.5; 0.7]$

$h_0 = \{0.5, 0.7\}$

$A_0^{\circ} = (d_0, h_0)$

$A_0^{\circ\circ} = (d_0, h_0)^{\circ} = d_{0x}^{\circ} = A_1 = \{g_1, g_2, g_3\}$

$d_1 = ([30; 35], [10; 12], [0.5; 0.8])$

$h_1 = \{0.5, 0.7, 0.8\}$

$A_0^{\circ\circ\circ} = A_1^{\circ} = (d_1, h_1)$

Description $d_0 = ([30; 35], [10; 12], [0.5; 0.7])$ is a 1/3-weak description with 1-allowed dropout, as soon as median from 0.5 and 0.7 equals 0.6.

The first stage of Query Based Regression Algorithm (QBRA) is mining α-weak premises with allowed ω-dropout, the second is to perform prediction for test object g_t based on the mined premises. Subsample size is a hyperparameter which is the number of objects being randomly extracted from *G*. Then α and ω hyperparameters are specified. They control for anti-support in terms of both frequency and magnitude. After objects $A_0 = \{g_1, \ldots, g_K\}$ are randomly extracted one calculates following pattern $d_0 = \delta(g_1) \sqcap \ldots \sqcap \delta(g_k) \sqcap \delta(g_{test})$ and density distribution function h_0 for target attribute values. If d_0 is an α - weak premise with allowed ω-dropout then it is added to the collection of premises that will be used for prediction. After premises mining, the next stage which is building up a prediction for target attribute based on mined premises. The resulting prediction was defined as a median of mixture of distributions from all premises.

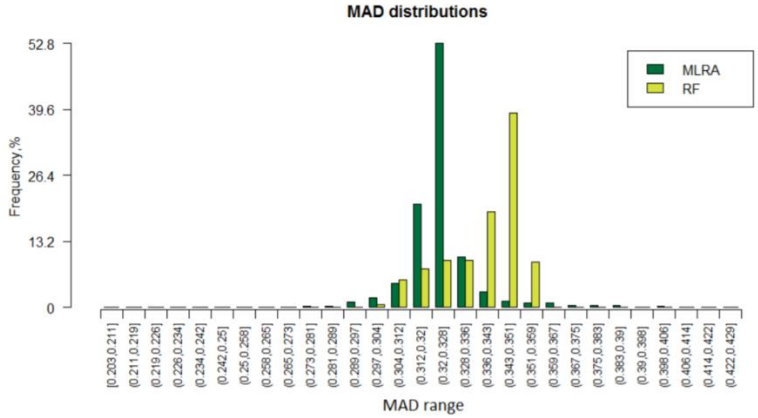
To test the algorithm, we used financial data from balance sheets and profit and loss statements of 612 corporate clients from the top-10 Russian bank. Among others factors we used assets-to-liabilities ratio, debt-to-equity ratio, earnings before taxes and interest payments, return on assets etc. These clients were assessed at the time of early insolvency signals and the resulting recovery rate was collected.

The accuracy of predictions was evaluated in terms of mean absolute deviation (MAD):

$$MAD = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

where y_i is a target attribute (recovery rate) for *i*-th client in the test set and $\hat{y}_i$ is predicted value. The algorithm was benchmarked with random forest model.

MAD distribution shows that lazy algorithm allows one to obtain prediction error relatively lower than the one with random forest tunings.



Distributions represent accuracy achieved for a large number of algorithm runs with unique combination of hyperparameter values.

Other benchmarks are provided below:

	MAD value
Naive model (median value for all test objects)	0.35
LinearRegression	0.30
DecisionTreeRegressor	0.28
Random Forest	0.24
AdaBoostRegressor	0.28
GradientBoostingRegressor	0.25

The **conclusion** summarizes and focuses on that the key feature of risk management practice is, regardless of the model accuracy, it must keep interpretability. Formal concept analysis offers attractive instruments to extract knowledge from data as soon as intents of concepts can be considered as associative rules. FCA-based algorithms are suitable for predictive modeling in areas where model interpretation clarity is of great priority. Also, the results show that these randomized modifications for classification and regression tasks outperform classical methods used in banks such as scorecards and decision trees in terms of Gini and mean absolute deviation. Therefore, it is argued that proposed FCA-based classification and regression algorithms can compete with ordinary statistical instruments adopted in banks and still provide the sets of rules which were relevant for loan applicants.

In **Appendix** programming code both for QBCA and MLRA is provided. Some key functions for meet operator, intent and extent calculation, premises mining and final predictions are provided. The language it is provided with is R (https://www.r-project.org/) as soon as it has intuitive syntax and vectorized language, so that the reader grasps the idea behind the algorithm realizations. However, for production implementations different languages are recommended such as Java or Spark (for distributed systems).

Results Summary

1. A randomized FCA-based algorithm for classification rules mining is developed.
2. The concept of α-weak premise and other parameters of the algorithm are introduced. Prediction accuracy analysis is performed with regard to algorithm hyperparameters values tuned on credit scoring data.
3. An algorithm is developed that allows to use the device of interval pattern structures in the problem of regression with several hyperparameters: the number of iterations, the alpha-threshold, the size of the subsample size, the omega - ω-dropout, penalty for a high variance value of the target variable on the right side of the expanded pattern (penalty for high deviation).
4. The concept of an *augmented interval pattern structure* is introduced, the concept of an ω-dropout for α-weak descriptions is introduced. New definitions help to solve regression problem via formal concepts analysis methods.
5. Proposed algorithms interpretability is analyzed from the standpoint of credit decision maker. The accuracy of the algorithms is compared with the models of credit scoring and other benchmarks (both “white-box” and “black-box”).
6. Query-based classification algorithm was developed with three hyperparameters: *number of iterations, alpha-threshold, subsample size*. Accuracy analysis of the algorithm predictions depending on the hyperparameters values was performed, an intuitive explanation is given for the results obtained.
7. The developed algorithms were implemented as program code in R language.