УДК 681.518.5

Д.А. Жуков

АНАЛИЗ КРИТЕРИЕВ КАЧЕСТВА КЛАССИФИКАЦИИ ПРИ ДИАГНОСТИКЕ ФУНКЦИОНИРОВАНИЯ ТЕХНИЧЕСКОГО ОБЪЕКТА[1]

Жуков Дмитрий Анатольевич, *окончил факультет информационных систем и технологий Ульяновского государственного технического университета, специалист по базе данных Ульяновского филиала конструкторского бюро ПАО «Туполев», аспирант кафедры «Прикладная математика и информатика» УлГТУ. Имеет научные труды в области статистических методов и машинного обучения. [e-mail: zh.dimka17@mail.ru].*

Аннотация

При прогнозировании исправности технического объекта по известным показателям его функционирования в процессе предшествующей эксплуатации решается задача бинарной классификации состояний объекта. Эта задача может быть решена методами машинного обучения, при этом наиболее объективной характеристикой качества диагностики является F-мера, что объясняется несбалансированностью классов: как правило, в выборке прецедентов, полученных по результатам эксплуатации объекта, количество данных об исправных состояниях значительно больше, чем о неисправных. Значения этой меры являются случайной величиной, поскольку оцениваются по контрольной выборке, формируемой случайным образом. Исследование показало, что распределение этой характеристики как для базовых, так и для агрегированных классификаторов близко к нормальному. На конкретном примере показано, что среднее значение F-меры при агрегировании превышает аналогичное значение, полученное с помощью базовых классификаторов.

Ключевые слова: техническая диагностика, исправность, показатели функционирования, машинное обучение, агрегированный подход, критерии качества классификации.

doi: 10.35752/1991-2927-2019-3-57-112-117

ANALYSIS OF CLASSIFICATION QUALITY CRITERIA FOR DIAGNOSTICS OF TECHNICAL OBJECT OPERATION

Dmitrii Anatolevich Zhukov, *graduated from the Faculty of Information System and Technologies of Ulyanovsk State Technical University; Database Specialist of the Ulyanovsk Branch of Tupolev Design Bureau, PJSC; Postgraduate Student at the Department of Applied Mathematics and Informatics of UlSTU; an author of publications in the field of statistical methods and machine learning. e-mail: zh.dimka17@mail.ru.*

Abstract

When predicting the technical object health based on the known indicators of its previous operation, the problem of binary classification of the object state is solved. This problem may be solved using the machine learning methods, and due to imbalance of classes, the most reliable diagnostics quality measure is the F-measure. As a rule, the healthy-state data amount contained in the set of the use cases based on the results of object operation exceeds the unhealthy state data amount. Values of this measure are random variables as they are estimated using validation dataset formed in a random manner. The study showed that the distribution of this measure both for basic classifiers and for aggregated classifiers is close to standard one. A specific example demonstrates that the average value of F-measure in aggregation process exceeds the similar value obtained using the basic classifiers.

Key words: technical diagnostics, healthy state, performance indexes, machine learning, aggregated approach, classification quality criteria.

1 Исследование выполнено при финансовой поддержке РФФИ и Правительства Ульяновской области, проект № 18-48-730001.

Введение

Диагностика состояния технического объекта в условиях эксплуатации проводится по результатам измерений косвенных показателей его функционирования. При этом могут использоваться методы машинного обучения. Решается задача бинарной классификации: необходимо построить модель, которая позволит оценить вероятность принадлежности нового состояния объекта к одному из двух классов – исправному или неисправному [1– 3].

Качество классификации оценивается по различным критериям: доле ошибок, F-мере, площади AUC под кривой ошибок и другим [4–7]. Для расчета исходная выборка разбивается случайным образом на два подмножества: обучающую и контрольную выборки. Первая используется для оценки параметров модели классификации, вторая – для оценки качества этой модели.

Поскольку разбиение выборки на обучающую и контрольную части производится случайным образом, значения критериев качества оказываются случайными величинами. Представляет интерес исследовать характер их распределения и стабильность при различных разбиениях.

Для повышения качества классификации предложено использование агрегированного подхода [8, 9]. При формировании единого решения об исправности объекта можно агрегировать результаты по среднему значению, по медиане и с помощью процедуры голосования. При этом агрегированные результаты учитывают особенности каждого из базовых методов. Необходимо оценить значимость изменения показателей качества при агрегировании.

Цель исследования – сравнительный анализ критериев качества бинарной классификации при диагностике функционирования технических объектов и обоснование выбора критерия, обеспечивающего объективную оценку качества распознавания исправности технического объекта при заданном наборе прецедентов.

1 Критерии качества диагностики

Для оценки качества бинарной классификации наиболее распространенным критерием является доля правильных ответов:

$$Accuracy = Q / N,$$

где Q – количество правильно классифицированных объектов контрольной выборки, а N – общий размер контрольной выборки. Чаще используется противоположная характеристика – доля (или процент) ошибок на контрольной выборке.

Иногда для оценки качества классификации применяют средний квадрат отклонений истинного класса в r-м наблюдении Y_r (0 или 1) от его прогнозируемого значения $\hat{Y}_r$ [10, 11]:

$$\sigma^2 = \frac{1}{N}\sum_{r=1}^{l}\left(Y_r - \hat{Y}_r\right)^2.$$

При несбалансированных классах (как правило, в выборке количество данных об исправных состояниях технического объекта значительно больше, чем о неисправных) доля ошибок не может объективно оценивать качество классификации [4, 5]. Более информативны точность $P = \dfrac{tp}{tp + fp}$ и полнота $R = \dfrac{tp}{tp + fn}$,

где tp – количество правильно классифицированных исправных состояний, fp – количество неправильно классифицированных исправных состояний, fn – количество неправильно классифицированных неисправных состояний.

На основе этих двух показателей может быть сформирован единый критерий

$F = \dfrac{2PR}{P + R}$ – это гармоническое среднее точности и полноты (F-мера): чем ближе значение F к единице, тем качество классификации выше.

В качестве еще одного функционала качества может быть выбрана площадь под ROC-кривой – кривой ошибок (receiver operating characteristics): AUC (area under the curve) [5–7]. ROC-кривая образуется, если по оси абсцисс отложить значения $fp(c)$, а по оси ординат – $tp(c)$, где c – заданный порог. Чем ближе значение AUC к единице, тем лучше качество диагностики. Значение AUC автоматически учитывает диспропорцию в представителях класса, а также имеет простую вероятностную интерпретацию: это вероятность того, что ответ на случайном объекте из класса 1 будет больше ответа на случайном объекте из класса 0. На рисунке 1 показаны такие кривые, построенные в системе Matlab для рассмотренного ниже примера диагностики, при использовании трех методов бинарной классификации: логистической регрессии, метода опорных векторов и наивного байесовского классификатора.

2 Статистический анализ критериев качества

Проводились наблюдения за системой водоочистки, функционирование которой определяется восемью показателями водоисточника (температура, цветность, мутность и др.), а исправность – показателями качества питьевой воды. Из полученных 527 наборов данных в 127 случаях система была признана неисправной (24%). Вначале определялись четыре рассмотренные критерия качества классификации по 11 базовым методам бинарной классификации, встроенным в библиотеку инструментов Statistics and Machine Learning Toolbox в пакете Matlab.

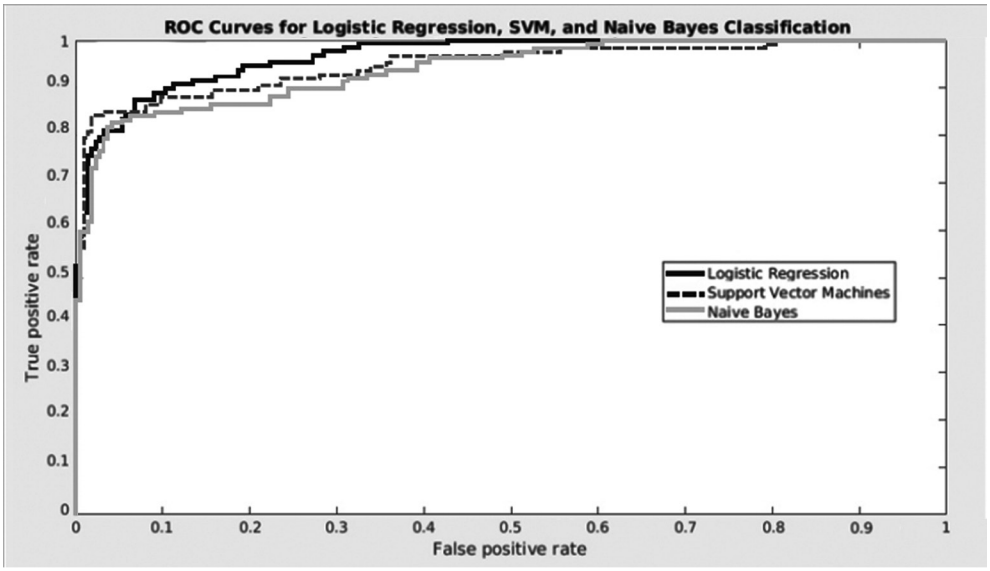

Рис. 1. ROC-кривые

	Метод	Ошибка по кросс-валидации	Средний квадрат отклонений	F-мера	AUC
1	ЛР	17.6778	38.2300	0.8862	0.9267
2	ДА	21.2627	4.3104	0.8593	0.6459
3	НС	22.5726	21.3320	0.8618	0.7968
4	AB	18.5994	21.3544	0.8808	0.8777
5	LB	17.0864	31.4935	0.8892	0.9119
6	GB	19.5356	18.9278	0.8731	0.8797
7	GrB	18.0334	21.2405	0.8885	0.9175
8	RB	19.9274	7.5670	0.8629	0.7317
9	БДР	14.8295	38.5689	0.9037	0.9427
10	МОВ	20.8636	30.5504	0.8586	0.7375
11	БК	20.2830	27.8099	0.8644	0.7914

Рис. 2. Критерии качества классификации

	Процент ошибок	Средний квадрат отклонений	F-мера	AUC
Процент ошибок	1			
Средний квадрат отклонений	0,02	1		
F-мера	−0,93	0,08	1	
AUC	−0,51	−0,53	0,38	1

Рис. 3. Матрица корреляций

На рисунке 2 показаны значения этих критериев для одного из вариантов разбиения исходной выборки на обучающую и контрольную части (ЛР – логистическая регрессия; ДА – дискриминантный анализ; БК – наивный байесовский классификатор; НС – нейронная сеть; МОВ – метод опорных векторов; БДР – бэггинг деревьев решений; GrB, AB, LB, GB, RB – различные варианты бустинга [8]). На рисунке 3 представлена соответствующая матрица корреляций, из которой видна сильная отрицательная корреляция между процентом ошибок и F-мерой; другие показатели коррелированны слабо или вообще некоррелированны.

Исследовался и характер распределения значений критериев качества, связанного со случайностью отбора прецедентов при формировании контрольной выборки. Испытания повторялись по 30 раз при одном и том же объеме контрольной выборки с использованием всех восьми показателей функционирования. На рисунке 4 показана гистограмма распределения значений F-меры с нанесенной кривой нормального распределения, построенная в системе Statistica для агрегирования по голосованию. Для проверки нормальности использован критерий Шапиро-Уилка, рекомендуемый при малых объемах выборки (до 50 наблюдений). Видно, что распределение можно считать нормальным на уровне значимости 0,05. Среднее значение F-меры при этом виде агрегирования оказалось 0,9015, а дисперсия – $1{,}81*10^{-5}$.

Аналогичные результаты получены и для других классификаторов (как базовых, так и агрегированных). Лучший результат по F-мере 0,9027 (наиболее близкий к единице) показало агрегирование по среднему значению.

3 Сравнительный анализ базовых и агрегированных классификаторов

При проведении испытаний лучший результат по F-мере показал БДР, для которого среднее значение F-меры составило 0,8937, лучший результат при агрегировании 0,9027, как отмечалось, показало агрегирование по среднему значению (АМС). Представляет интерес оценить значимость этого расхождения: более высокое значение по АМС по сравнению с БДР получено случайным образом, или это закономерный результат.

Учитывая, что распределения F-меры и для АМС, и для БДР близко к нормальному, можно воспользоваться стандартными алгоритмами проверки гипотез, например, с помощью электронных таблиц Excel [12].

Вначале проверяем гипотезу о равенстве дисперсий

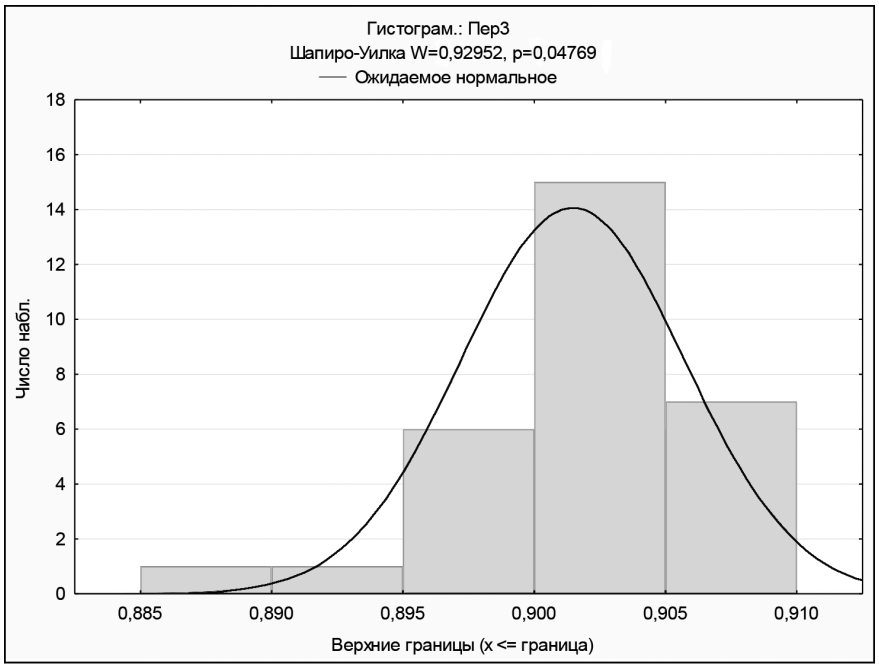

Рис. 4. Распределение F-меры

Двухвыборочный F-тест для дисперсии		
	АМС	*БДР*
Среднее	0,902663	0,893676667
Дисперсия	1,59E−05	1,4515E−05
Наблюдения	30	30
df	29	29
F	1,093592	
$P(F<=f)$ одностороннее	0,405633	
F критическое одностороннее	1,860811	

а)

Двухвыборочный t-тест с одинаковыми дисперсиями		
	АМС	*БДР*
Среднее	0,902663	0,893677
Дисперсия	1,59E−05	1,45E−05
Наблюдения	30	30
Объединенная дисперсия	1,52E−05	
Гипотетическая разность средних	0	
df	58	
t-статистика	8,929053	
$P(T<=t)$ одностороннее	8,68E-13	
t критическое одностороннее	1,671553	
$P(T<=t)$ двухстороннее	1,74E−12	
t критическое двухстороннее	2,001717	

б)

Рис. 5. Проверка гипотезы о равенстве значений F-меры для базовых и агрегированных классификаторов

с помощью теста Фишера (рис. 5а). Видим, что выборочное значение статистики 1,09 меньше критического 1,86; статистика попадает в область принятия решения: гипотеза о равенстве дисперсий принимается. На этом основании для проверки гипотезы о равенстве средних значений используем t-тест с одинаковыми дисперсиями (рис. 5б).

Проверяется нулевая гипотеза о равенстве средних в двух рассматриваемых совокупностях при альтернативе: среднее при агрегировании АМС больше среднего при лучшем из базовых классификаторов БДР. Используется односторонний критерий. Выборочное значение t-статистики 8,93 больше критического одностороннего значения 1,67 (как и двухстороннего 2,00, но в данном случае это несущественно), то есть попадает в критическую область: нулевая гипотеза о равенстве средних отвергается. Таким образом, среднее значение F-меры

при агрегировании значимо больше, чем среднее при БДР.

4 Анализ структуры агрегированных классификаторов

АМС: НС + БДР + АВ + GB	0,9048
АМС: БДР + LB	0,9055
АМС: БДР + LB	0,9003
АМС: GrB + БДР + LB	0,904
АМС: НС + БДР + LB + GB	0,9075
АМС: БДР + GB	0,9027
АМС: GrB + БДР + LB + GB	0,9035
АМС: МОВ + ЛР + БДР + LB + GB	0,9015
АМС: НС + БДР + GB	0,9011
АМС: НС + GrB + БДР + LB + GB	0,9028
АМС: БДР + GB	0,9061

Рис. 6. Структуры агрегированных классификаторов

На рисунке 6 показана часть выборки при агрегировании по среднему значению АМС: в левом столбце – структура агрегата, в правом – значение F-меры. Например, запись в первой строке АМС: НС + БДР + АВ + GB означает, что в данном опыте наилучшим вариантом агрегирования по среднему значению при использовании в качестве критерия качества диагностики F-меры оказалась совокупность из четырех базовых классификаторов, включающая НС, БДР и методы бустинга AdaBoost и GentleBoost. Количество базовых классификаторов на рисунке 6 колеблется от двух до пяти, а в общем случае может достигать и одиннадцати.

Вместе с тем, рисунок 6 показывает, что значения F-меры различаются не слишком существенно. Имеет смысл проверить гипотезу о том, что увеличение в структуре агрегата числа базовых методов больше двух несущественно влияет на значения F-меры. С этой целью разобьем всю полученную выборку из 90 наблюдений (три метода агрегирования по 30 опытов в каждом) на два подмножества. В первое подмножество включим данные по агрегатам, состоящим только из двух компонент (таких оказалось 48 из 90), во второе – все остальные значения.

Проверка гипотезы о равенстве средних в этих подмножествах, проведенная по аналогии с методом, иллюстрированном на рисунке 5, показывает справедливость этой гипотезы: среднее значение F-меры не изменяется при увеличении количества базовых классификаторов в структуре агрегата.

Отсюда вытекает важный вывод о возможности резкого сокращения времени на вычисления. Вместо перебора всех вариантов агрегирования для поиска максимального значения F-меры (а это 3*(211– 1) = 6141 вариант! [8]) достаточно перебрать только варианты, включающие по два базовых метода (3*11!/2!9! = 165).

Можно учесть и еще одно обстоятельство. Во всех 90 опытах в состав агрегата вошел лучший из базовых методов для рассматриваемого объекта – БДР. Учет этого факта позволяет сократить количество перебираемых вариантов до 30.

Следует, однако, иметь в виду, что представленные результаты получены на эксперименте лишь с одним техническим объектом – системой водоочистки. Тем не менее, этот опыт показывает, что предложенный подход имеет смысл апробировать и для диагностики любой другой исследуемой системы. Исследования, проведенные с двумя другими объектами (при вибромониторинге гидроагрегата и анализе исправности счетчиков горячего водоснабжения) подтвердили сформулированные результаты.

Заключение

Проведенный анализ критериев качества бинарной классификации при диагностике функционирования технических объектов показал, что наиболее объективной характеристикой является F-мера. Это объясняется несбалансированностью классов: как правило, в выборке прецедентов, полученных по результатам эксплуатации объекта, количество данных об исправных состояниях значительно больше, чем о неисправных. Как и для любой другой характеристики качества диагностики, значения этой меры являются случайной величиной, поскольку оцениваются по контрольной выборке, формируемой случайным образом. Впервые проведено исследование свойств этой характеристики, которое показало, что ее распределение как для базовых, так и для агрегированных классификаторов, близко к нормальному. На конкретном примере показано, что среднее значение F-меры при агрегировании превышает аналогичное значение, полученное с помощью базовых классификаторов. При построении агрегированного классификатора достаточно рассмотреть комбинацию из двух базовых методов: более сложные комбинации в рассмотренном примере не выявили улучшения качества диагностики.

СПИСОК ЛИТЕРАТУРЫ

1. Жуков Д.А., Клячкин В.Н. Диагностика исправности технического объекта с использованием пакета Matlab // Перспективные информационные технологии: тр. Междунар. науч.-техн. конф. – Самара : Издательство Самарского научного центра РАН, 2018. – С. 55–57.

2. Жуков Д.А., Клячкин В.Н. Влияние объема контрольной выборки на качество диагностики состояния технического объекта // Автоматизация процессов управления. – 2018. – № 2 (52). – С. 90–95.

3. Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Влияние способа отбора значимых показателей на качество диагностики состояния технического объекта // Автоматизация. Современные технологии. – 2019. – Т. 73, № 1. – С. 32–36.

4. Соколов Е.А. ФКН ВШЭ. Лекция 4. Линейная классификация. – URL: https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf (дата обращения: 01.03.2019).

5. Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves // Proceedings of the 23rd international conference on Machine learning. Pittsburgh, 2006. pp. 233–240.

6. Дьяконов А.М. AUC ROC (площадь под кривой ошибок). – URL: https:// dyakonov.org/2017/07/28/auc-roc-площадь-под-кривой-ошибок/#more-5362 (дата обращения: 01.03.2019).

7. Жуков Д.А., Клячкин В.Н. Критерии качества диагностики функционирования технических объектов методами машинного обучения // Информатика, моделирование, автоматизация проектирования. X Всероссийская школа-семинар аспирантов, студентов и молодых ученых : сб. науч. тр. – Ульяновск : УлГТУ, 2018. – С. 87–90.

8. Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Диагностика технического состояния аппаратуры с использованием агрегированных классификаторов // Радиотехника. – 2018. – № 6. – С. 46–49.

9. Klyachkin V.N., KuvayskovaYu.E., Zhukov D.A. The use of aggregate classifiers in technical diagnostics, based on machine learning // CEUR Workshop Proceedings. – Data Science. Information Technology and Nanotechnology. 2017. Vol. 1903. pp. 32–35.

10. Шунина Ю.С., Алексеева В.А., Клячкин В.Н. Прогнозирование кредитоспособности клиентов банка на основе методов машинного обучения // Финансы и кредит. – 2015. – № 27 (651). – С. 2–12.

11. Клячкин В.Н., Шунина Ю.С. Система оценки кредитоспособности заемщиков и прогнозирования возврата кредитов // Вестник компьютерных и информационных технологий. – 2015. – № 11 (137). – С. 45–51.

12. Клячкин В.Н., Кувайскова Ю.Е., Алексеева В.А. Статистические методы анализа данных. – М. : Финансы и статистика, 2016. – 240 с.

REFERENCES

1. Zhukov D.A., Kliachkin V.N. Diagnostika ispravnosti tekhnicheskogo obekta s ispolzovaniem paketa Matlab [Diagnostics of Serviceability of a Technical Object with Matlab Package]. *Perspektivnye informatsionnye tekhnologii: tr. Mezhdunar. nauch.-tekhn. konf.* [Proc. of Int. Sci. and Tech. Conf. Advanced Information Technologies]. Samara, Samara Research Center of RAN Publ., 2018, pp. 55–57.

2. Zhukov D.A., Kliachkin V.N. Vliianie obema kontrolnoi vyborki na kachestvo diagnostiki sostoianiia tekhnicheskogo obekta [The Effect of the Control Sample Volume on the Quality of Diagnostics of the Technical Object State]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2018, no. 2 (52), pp. 90–95.

3. Kliachkin V.N., Kuvaiskova Iu.E., Zhukov D.A. Vliianie sposoba otbora znachimykh pokazatelei na kachestvo diagnostiki sostoianiia tekhnicheskogo obekta [The Influence of the Significant Indicators Selection Method on the Quality of the Technical Object Diagnostics]. *Avtomatizatsiia. Sovremennye tekhnologii* [Automation. Modern Technologies], 2019, vol. 73, no. 1, pp. 32–36.

4. Sokolov E.A. *FKN VShE. Lektsiia 4. Lineinaia klassifikatsiia* [Linear Classification. Lecture 4. Faculty of Computer Science at the Higher School of Economics]. Available at: https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf (accessed: 01.03.2019).

5. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning.* Pittsburgh, 2006, pp. 233–240.

6. Diakonov A.M. AUC ROC (*ploshchad pod krivoi oshibok*) [AUC ROC (Area Under Curve Receiver Operating Characteristic]. Available at: https:// dyakonov.org/2017/07/28/auc-roc-ploshchad-pod-krivoy-oshibok/#more-5362 (accessed: 01.03.2019).

7. Zhukov D.A., Klachkin V.N. Kriterii kachestva diagnostiki funktsionirovaniia tekhnicheskikh obektov metodami mashinnogo obucheniia [Quality Criteria for Diagnosing the Functioning of Technical Objects Using Machine Learning Methods]. *Informatika, modelirovanie, avtomatizatsiia proektirovaniia. X Vserossiiskaia shkola-seminar aspirantov, studentov i molodykh uchenykh. Sb. nauch. tr.* [Informatics, Modeling, Computer-Aided Design. Proc. of the 10th Russian Workshop for Postgraduates, Students, and Young Scientists]. Ulyanovsk, UlSTU Publ., 2018, pp. 87–90.

8. Kliachkin V.N., Kuvaiskova Iu.E., Zhukov D.A. Diagnostika tekhnicheskogo sostoianiia apparatury s ispolizovaniem agregirovannykh klassifikatorov [Diagnostics of Technical State of the Equipment Using Aggregated Classifiers]. *Radiotekhnika* [Journal Radioengineering], 2018, no. 6, pp. 46–49.

9. Kliachkin V.N., Kuvaiskova Iu.E., Zhukov D.A. The Use of Aggregate Classifiers in Technical Diagnostics, Based on Machine Learning. *CEUR Workshop Proceedings. Data Science. Information Technology and Nanotechnology.* 2017, vol. 1903, pp. 32–35.

10. Shunina Iu.S., Alekseeva V.A., Kliachkin V.N. Prognozirovanie kreditosposobnosti klientov banka na osnove metodov mashinnogo obucheniia [Forecasting the Customers' Creditworthiness through Machine Learning Methods]. *Finansy i kredit* [Finance and Credit], 2015, no. 27 (651), pp. 2–12.

11. Kliachkin V.N., Shunina Iu.S. Sistema otsenki kreditosposobnosti zaemshchikov i prognozirovaniia vozvrata kreditov [System for Borrowers' Creditworthiness Assessment and Repayment of Loans Forecasting]. *Vestnik kompiuternykh i informatsionnykh tekhnologii* [Herald of Computer and Information Technologies], 2015, no. 11 (137), pp. 45–51.

12. Kliachkin V.N., Kuvaiskova Iu.E., Alekseeva V.A. *Statisticheskie metody analiza dannykh* [Statistical Methods for Data Analysis]. Moscow, Finansy i statistika Publ., 2016. 240 p.