Conclusion. *In the paper, the phase equilibrium curves of binary FCC alloys with limited solubility of components were investigated by computer simulations by lattice Monte Carlo methods using the diffusion couple method. The Glauber and Metropolis algorithms based on exchange and vacancy diffusion mechanisms, as well as Residence Time Algorithm for the vacancy mechanism were used for the numerical simulation. The phase diagram of FCC binary system with limited solubility of components was constructed by lattice Monte Carlo algorithms based on the simulation of kinetics of diffusion processes at an atomic level. The existence of a single curve of phase equilibrium of the modeled binary system is confirmed for different lattice Monte Carlo methods. The linear dependences of the difference of the reduced temperatures of the phase equilibrium curves on the concentration for a model of a regular solid solution and a model binary system based on the lattice Monte Carlo method were obtained.*

Keywords: *Monte Carlo, binary alloy, regular solid solution, diffusion couple method.*

Одержано редакцією 08.08.2019 р.
Прийнято до публікації 09.10.2019 р.

УДК 004.85:519.6

DOI 10.31651/2076-5886-2019-2-86-95

PACS 02.70.Wz, 07.05.Kf, 07.05.Mh, 07.05.Tp

ПІСКУН Олександр Варфоломійович
кандидат технічних наук, доцент,
завідувач кафедри прикладної математики та інформатики, Черкаський національний університет імені Богдана Хмельницького
e-mail: piskun@ukr.net
ORCID 0000-0001-5334-6337

АНАЛІЗ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ЗАДАЧІ БІНАРНОЇ КЛАСИФІКАЦІЇ

У роботі був проведений аналіз існуючих найбільш поширених методів класифікації на предмет їх використання в задачі діагностики серцевих захворювань. Розглянуто основні метрики якості моделей бінарної класифікації, які можуть бути використані при ухваленні рішення про оптимальність розробленої моделі. Дослідження моделей проводились без і з оптимізацією параметрів. Оптимізація параметрів моделей проведена, використовуючи криві валідації з подальшим пошуком по сітці з крос-валідацією кожної комбінації параметрів. Найкращі результати показали методи DecisionTreeClassifier, GradientBoostingClassifier та GaussianNB.

Ключові слова: *машинне навчання, метрики якості, бінарна класифікація, алгоритми класифікації*

Вступ

Бінарна класифікація - одна з найбільш поширених проблем прикладної статистики та машинного навчання, яка вирішується в багатьох прикладних областях - в медицині, біології, метеорології, аналізі поштових повідомлень, класифікації текстів, зображень і т.д.

Розглянемо задачу бінарної класифікації об'єктів, в якій кожен об'єкт K_i (i = 1, ..., N) характеризується m-мірним вектором ознак (X_1 ... X_m). Ці характеристики (або ознаки) можуть приймати як числові, так і нечислові значення та утворюють вибірку для подальших досліджень. Потрібно на підставі значень ознак передбачити вихідну характеристику об'єктів y, яка може приймати одне з двох значень (0 або 1).

Існує багато методів класифікації, такі як дерева рішень, нейронні мережі, байесовский класифікатор, метод опорних векторів, логістична регресія і ін., які використовують різний математичний апарат і різні підходи при реалізації [1]. Однак, їх ефективність залежить від конкретної задачі і на сьогоднішній день не існує методів,

які могли б однозначно ефективно вирішити задачу класифікації. Тому, доцільно проводити апробацію декількох методів бінарної класифікації та знаходити оптимальні для вирішення поставленої задачі.

Метою статті є знаходження найкращого алгоритму машинного навчання для вирішення задачі класифікації щодо визначення наявності або відсутності захворювань серця у людини на прикладі реальних даних.

Виклад основного матеріалу

Оцінка якості моделей класифікації є важливим аспектом у всіх областях, для яких розробляються моделі машинного навчання. Дана оцінка якості відповідає на питання, наскільки добре отриманий класифікатор розділяє класи, що нас цікавлять, на деякій вибірці.

Метрики якості моделей бінарної класифікації [2-4]

Стосовно багатокласової класифікації, метрика відображає точність класифікації для усіх класів разом узятих або для кожного окремого класу.

Accuracy. Нехай класифікатор видає мітку класу. Позначимо: y_i – мітка і-го об'єкту, $\hat{y}_t$ – відповідь на цьому об'єкті нашого алгоритму, m – число об'єктів у вибірці, тоді частка об'єктів, по яким класифікатор прийняв правильне рішення, визначається як:

$$Accuracy(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{m} I[\hat{y}_i = y_i].$$

Частка вірних відповідей не враховує ціни помилок.

Confusion matrix. Бінарна класифікація використовується в задачах, де об'єкти вибірки діляться на два класи – позитивні (positive) та негативні (negative). У свою чергу, сама модель бінарної класифікації привласнює об'єктам також дві мітки – positive або negative (рис. 1а). А оскільки модель буде працювати з помилками щодо тестової вибірки, то в результаті бінарної класифікації всі об'єкти вибірки розбиваються на чотири типи, утворюючи матрицю неточностей / помилок (confusion matrix) (рис. 1б):

1) істинно-позитивні (true positive – TP);
2) істинно-негативні (true negative – TN);
3) помилково-позитивні (false positive – FP) – помилка 1-го роду (type I error);
4) помилково-негативні (false negative – FN) – помилка 2-го роду (type II error).

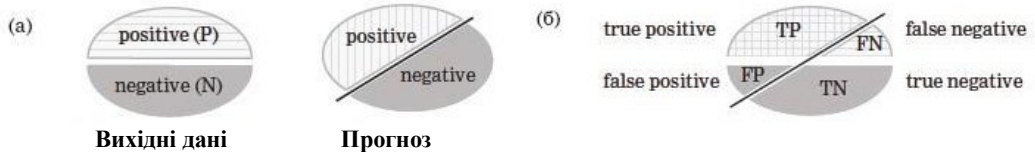

Рис. 1. Вихідні та прогнозні дані бінарної класифікації (а)
та матриця неточностей (б) [4]

Матриця неточностей бінарної класифікації - матриця розміру 2×2, ij-я позиція якої дорівнює числу об'єктів і-го класу, яким алгоритм присвоїв мітку j-го класу (рис. 2).

		Predicted Labels: Class 0	Predicted Labels: Class1
Real Labels	Class 0	True Negatives (TN)	False Positives (FP)
	Class1	False Negatives (FN)	True Positives (TP)

Рис. 2. Матриця неточностей для бінарної класифікації

Використовуючи наведену вище матрицю неточностей можна отримати кілька метрик якості моделі бінарної класифікації, які при цьому не є взаємовиключними, доповнюють одна одну та можуть бути використані в процесі прийняття рішення про оптимальну модель в кожному конкретному випадку. Наприклад, у медичній сфері помилка 1-го роду є найбільш критичною, тому що може бути краще поставити більш песимістичний діагноз, ніж більш оптимістичний [5].

Precision і Recall. Для оцінки якості роботи алгоритму на кожному з класів окремо вводяться метрики Precision (точність) і Recall (повнота):

$$\text{Precision} = \frac{TP}{TP + FP}, \ \text{Recall} = \frac{TP}{TP + FN}.$$

Точність - це частка об’єктів, які дійсно належать позитивному класу щодо всіх об’єктів, віднесених класифікатором до цього класу.

Повнота - це частка знайдених класифікатором об’єктів, які належать позитивному класу щодо всіх об’єктів цього класу в тестовій вибірці.

F-міра. Зрозуміло, що чим вище точність і повнота, тим краще. Однак, максимальна точність та повнота недосяжні одночасно і доводиться шукати деякий баланс. F-міра об'єднує в собі інформацію про точність та повноту алгоритму.

Міра F_1 є гармонійне середнє між точністю та повнотою:

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

При цьому, якщо ми хочемо віддати перевагу або точності, або повноті, можна скористатися розширеною версією F-міри, що має параметр β, який можна використовувати для балансування точності та повноти:

$$F_\beta = \left(1 + \beta^2\right)\frac{\text{Precision} \times \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}.$$

Опис набору даних

Для дослідження використовується набір даних (датасет) з серцевих захворювань Statlog (Heart) [6]. Дані містять 13 атрибутів (табл. 1) і 270 спостережень.

Таблиця 1

Атрибути даних з серцевих захворювань

№	Назва	Опис
1	age	вік пацієнта в роках
2	sex	стать
3	chest_pain_type	тип болю в грудях
4	resting_blood_pressure	артеріальний тиск у спокої (мм рт. ст.)
5	serum_cholestoral_mg_per_dl	рівень холестерину в крові в мг/дл
6	fasting_blood_sugar_gt_120_mg_per_dl	рівень цукру в крові > 120 мг/дл
7	resting_ekg_results	ЕКГ результати в стані спокою
8	max_heart_rate_achieved	максимальна частота серцевих скорочень
9	exercise_induced_angina	стенокардія, викликана фізичними вправами
10	oldpeak_eq_st_depression	депресія сегменту ST на ЕКГ
11	slope_of_peak_exercise_st_segment	нахил пікового сегменту ST
12	num_major_vessels	кількість судин, забарвлених при флюороскопії
13	thal	тип дефекту

Змінна для прогнозування **y** може приймати два значення: 0 – відсутність, 1 – наявність серцевих захворювань.

Дослідження проводимо з використанням бібліотеки машинного навчання Scikit-Learn, в середовищі Python 3.

Аналіз та підготовка даних

Завантажуємо дані та перевіряємо їх на наявність пропущених значень:

```
In [125]: X.isnull().sum()
Out[125]:
slope_of_peak_exercise_st_segment       0
thal                                    0
resting_blood_pressure                  0
chest_pain_type                         0
num_major_vessels                       0
fasting_blood_sugar_gt_120_mg_per_dl    0
resting_ekg_results                     0
serum_cholesterol_mg_per_dl             0
oldpeak_eq_st_depression                0
sex                                     0
age                                     0

max_heart_rate_achieved                 0
exercise_induced_angina                 0
dtype: int64
```

Пропусків немає.

Проводимо перевірку наявності дисбалансу цільових класів. Результати представлені на рис. 3.

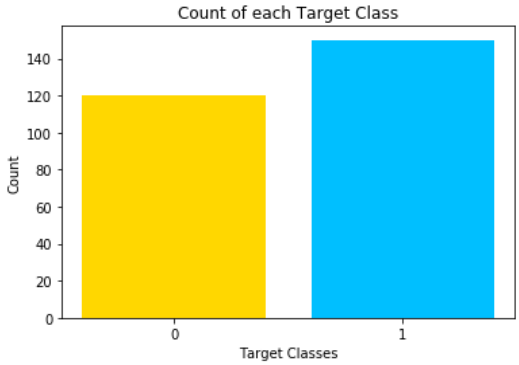

Рис. 3. Кількість об’єктів цільових класів

Класи можна вважати збалансованими (об’єктів класу «0» - 120, «1» - 150). Побудуємо гістограми атрибутів (рис. 4).

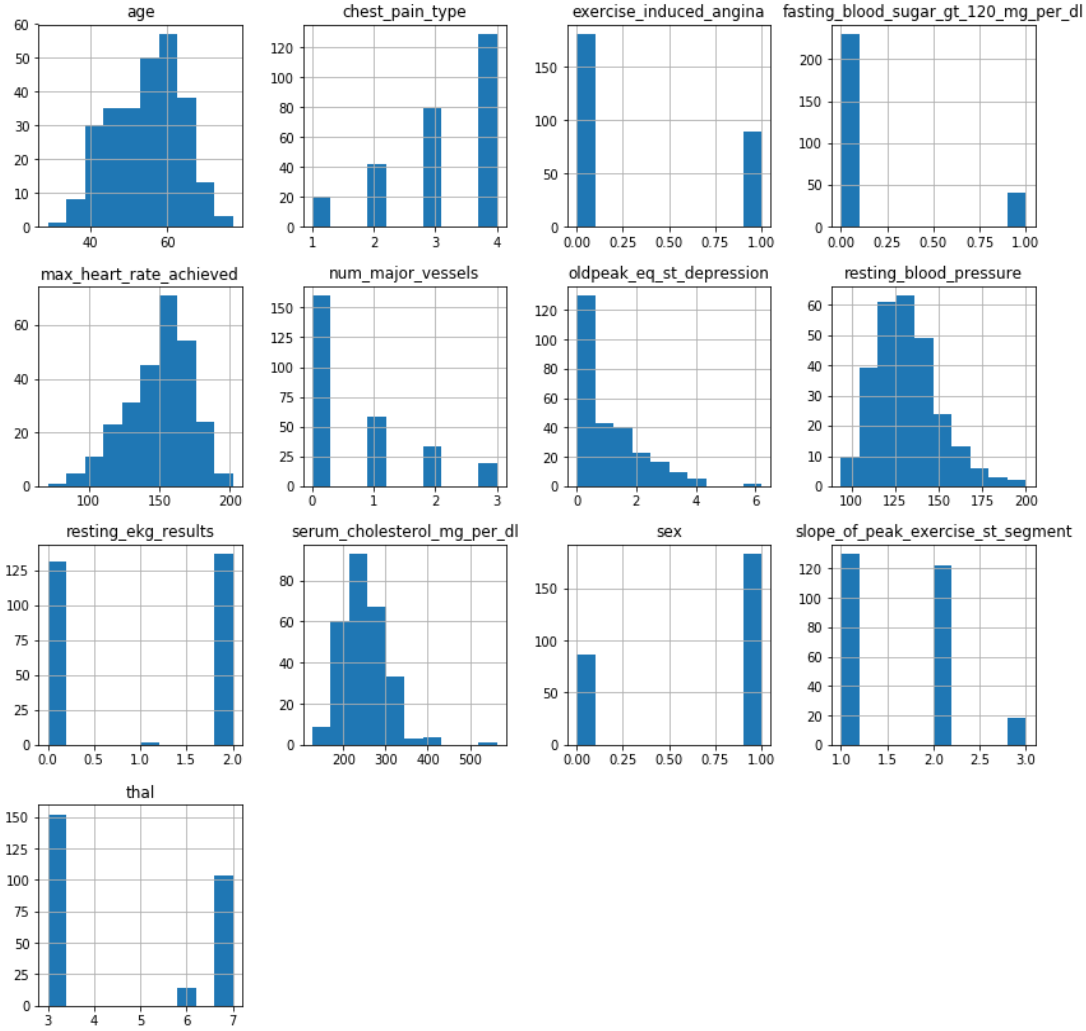

Рис. 4. Гістограми атрибутів

Як видно з гістограм, наш датасета включає в себе як кількісні, так і категоріальні дані, що вимагає їх попередньої обробки: категоріальні перевести в бінарні, а кількісні - нормалізувати.

Побудуємо матрицю кореляцій атрибутів (рис.5).

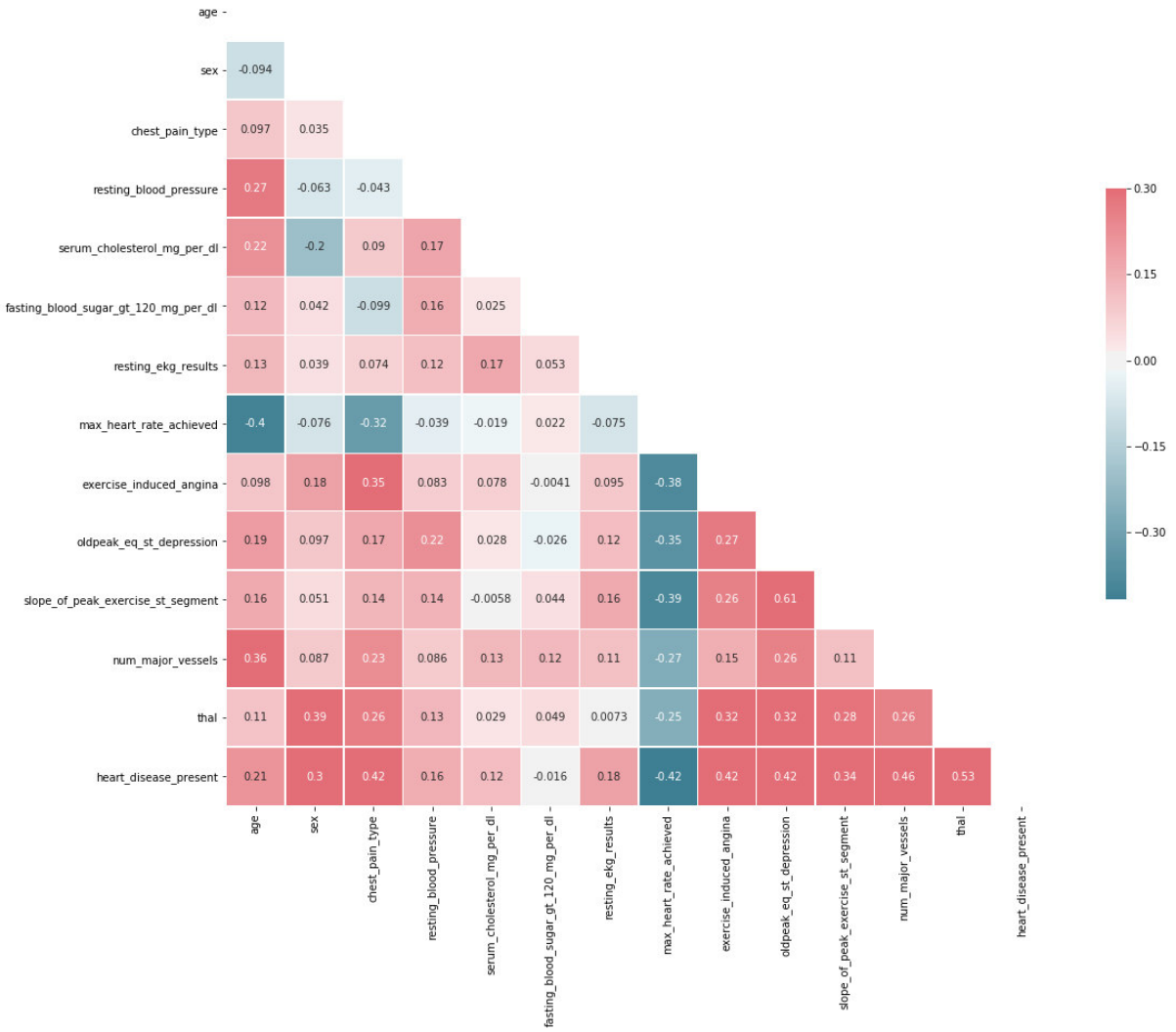

Рис. 5. Кореляційна матриця атрибутів

Як видно з матриці кореляцій, атрибути не мають між собою достатньо високих кореляцій, тому їх можна вважати незалежними.

Виконаємо попередню обробку даних: перетворимо категоріальні дані у бінарні, нормалізуємо усі числові до одного діапазону [0..1].

Для подальшого аналізу даних проведемо візуалізацію змінної **y** за допомогою методу t-SNE. T-Distributed Stochastic Neighbor Embedding (t-SNE) - це метод нелінійного зменшення розмірності, який добре підходить для візуалізації багатовимірних наборів даних. Метод моделює кожен об'єкт простору високої розмірності дво- або тривимірною точкою таким чином, що близькі за характеристиками елементи даних в багатовимірному просторі (наприклад, датасета з великим числом стовпців) проектуються в сусідні точки, а різнорідні об'єкти з великою ймовірністю моделюються точками, які стоять далеко одна від одної [7]. Результати представлені на рис. 6.

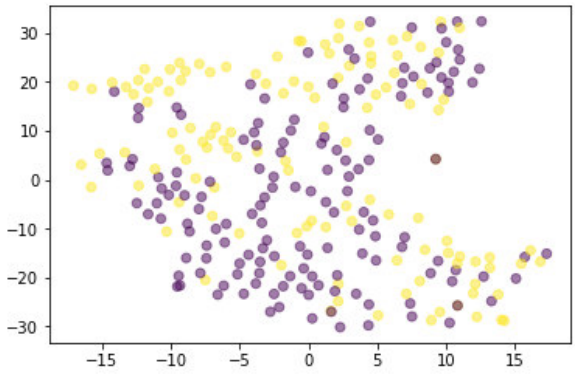

Рис. 6. T-SNE візуалізація цільових классів

Отримані результати показують, що ми маємо лінійно нероздільну вибірку.

Цільові класи не вдасться просто лінійно розділити, тому будемо застосовувати різні методи класифікації з подальшим порівнянням за допомогою відповідної метрики оцінки моделі.

Розділяємо вибірку на тренувальну та тестову (75% / 25%).

Застосуємо набір загально використовуваних класифікаторів [8] з параметрами по замовченню:

names = ['Logit', "Nearest Neighbors", "RBF SVM", "Gaussian Process", "Decision Tree", "Random Forest", "Multi-layer Perceptron", 'GradBoost', 'StochGradBoost', "AdaBoost", "Naive Bayes"]

```
classifiers = [
    LogisticRegression(),
    KNeighborsClassifier(),
    SVC(kernel="rbf", probability=True),
    GaussianProcessClassifier(1.0 * RBF(1.0)),
    DecisionTreeClassifier(),
    RandomForestClassifier(),
    MLPClassifier(),
    GradientBoostingClassifier(),
    SGDClassifier(),
    AdaBoostClassifier(),
    GaussianNB()]
```

Recall є для нас основною метрикою, так як невиявлення хвороби при її наявності може призвести до смерті пацієнта. Якщо ж діагноз буде поставлено здоровій людині, то додаткове обстеження дасть можливість виявити помилку без загрози життю людини. Тому далі ми будемо використовувати accuracy, як метрику для оцінки ефективності моделі загалом, та recall – при вирішенні поставленої задачі.

Розраховуємо точність і повноту на тренувальному та тестовому наборах з крос-валідацією по 10 вікнах. Таким чином, ми отримаємо середнє значення точності і повноти моделі та їх стандартні відхилення, що дасть уявлення про стабільну ефективність моделі і знизить вплив вибірки на результат.

Отримані результати точності та повноти для моделей преставлені в табл. 2.

Значення точності і повноти недозволено низькі для нашої задачі та до того ж мають великий розкид значень. Recall на тестовій вибірці є нижчим за 80%, тобто більше 200 з 1000 хворих пацієнтів віднесені до здорових, тільки DecisionTree Classifier і GaussianNB дали результати 84% і 93% відповідно, але при розкиді значень у 16% і

13%. Це говорить про істотну залежність результатів від вибірки та вимагає підвищення стабільності роботи моделей.

Таблиця 2

Значення метрик точності та повноти для моделей моделей з параметрами по замовченню на тренувальній та тестовій виборках

Classifier (Default parameters)	**Accuracy Train**	**Acc Tr Std**	**Accuracy Test**	**Acc Test Std**	**Recall Train**	**Recall Tr Std**	**Recall Test**	**Recall Test Std**
LogisticRegression	0,85	0,05	0,78	0,09	0,80	0,10	0,74	0,20
KNeighbors Classifier	0,85	0,06	0,75	0,17	0,80	0,14	0,64	0,24
SVC	0,85	0,05	0,78	0,13	0,79	0,13	0,74	0,20
GaussianProcess Classifier	0,87	0,05	0,76	0,10	0,81	0,11	0,71	0,23
DecisionTree Classifier	0,76	0,05	0,81	0,13	0,78	0,14	0,84	0,16
RandomForest Classifier	0,82	0,04	0,75	0,11	0,74	0,12	0,64	0,18
MLPClassifier	0,86	0,07	0,75	0,11	0,83	0,09	0,74	0,20
GradientBoosting Classifier	0,82	0,07	0,79	0,10	0,80	0,12	0,71	0,23
SGDClassifier	0,78	0,05	0,69	0,13	0,70	0,17	0,67	0,26
AdaBoostClassifier	0,79	0,07	0,72	0,21	0,77	0,10	0,77	0,21
GaussianNB	0,80	0,08	0,63	0,15	0,61	0,20	0,93	0,13

Виконаємо оптимізацію параметрів для усіх застосованих моделей. Оптимізація проводиться, використовуючи криві валідації з подальшим пошуком по сітці з крос-валідацією кожної комбінації параметрів [9]. Результати представлені в табл. 3.

Таблиця 3

Значення метрик точності та повноти для моделей з оптімізованими параметрами на тренувальній та тестовій виборках

Classifier (Best parameters)	**Accuracy Train**	**Acc Tr Std**	**Accuracy Test**	**Acc Test Std**	**Recall Train**	**Recall Tr Std**	**Recall Test**	**Recall Test Std**
LogisticRegression	0,86	0,05	0,79	0,07	0,82	0,11	0,78	0,15
KNeighbors Classifier	0,86	0,06	0,76	0,16	0,81	0,11	0,67	0,26
SVC	0,87	0,05	0,78	0,09	0,83	0,11	0,74	0,20
GaussianProcess Classifier	0,85	0,07	0,79	0,10	0,78	0,16	0,71	0,18
DecisionTree Classifier	0,76	0,05	0,81	0,13	0,78	0,14	0,84	0,16
RandomForest Classifier	0,82	0,05	0,82	0,09	0,79	0,13	0,80	0,16
MLPClassifier	0,85	0,05	0,72	0,10	0,80	0,12	0,78	0,21
GradientBoosting Classifier	0,85	0,06	0,82	0,13	0,84	0,09	0,87	0,16
SGDClassifier	0,83	0,04	0,75	0,09	0,82	0,17	0,80	0,22
AdaBoostClassifier	0,82	0,06	0,82	0,15	0,82	0,09	0,83	0,22
GaussianNB	0,81	0,08	0,64	0,16	0,64	0,21	0,93	0,13

Як видно з табл. 3, оптимізація параметрів не призвела до суттєвого покращення якості моделей, за винятком GradientBoostingClassifier, яка посіла друге місце після GaussianNB. Так як ми маємо сбалансований набір даних без викидів, вибірка, швидше за все, недостатньо велика для забезпечення оптимального тренування моделей.

Таким чином, ми можемо виділити три моделі, які найбільш підходять для даної конкретної задачі - DecisionTree Classifier, GradientBoosting Classifier та GaussianNB з показниками recall 0,84; 0,87 та 0,93 відповідно. Стандартні відхилення 0,16; 0,16 та 0,13 вказують на чутлівість моделей до вибірки даних.

Висновки

В роботі розглянуто та вибрано метрики якості моделей бінарної класифікації з урахуванням поставленого завдання. Для оцінки ефективності моделі загалом використовувалась метрика accuracy, при вирішенні поставленої задачі – recal.

Досліджено 11 найбільш поширених методів класифікації на предмет їх використання для вирішення задачі класифікації щодо визначення наявності або відсутності захворювань серця у людини на прикладі реальних даних. Отримано показники якості роботи моделей (з кросс-валидацією по 10 вікнам) з параметрами по замовченню та з оптимізацією параметрів. Найкращі результати показали методи DecisionTreeClassifier, GradientBoostingClassifier та GaussianNB, але вони не можуть бути визнані достатніми в медичній практиці.

Майбутніми дослідженнями будуть детальний розгляд моделей DecisionTreeClassifier, GradientBoostingClassifier та GaussianNB, зокрема, знаходження причин розкиду значень метрики recall для його зниження і, відповідно, підвищення стабільності точності моделей. На основі даних трьох методів буде побудована асамблейна класифікація розпізнавання хвороби серця, що може підвищити точність моделі.

Список використаної літератури:

1. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? / M.F. Delgado, E. Cernadas, S. Barro, D. Amorim // Journal of Machine Learning Research. – 2014. – V. 15. – P. 3133-3181.
2. Лабинцев, Е. Метрики в задачах машинного обучения [Електронний ресурс] / Е. Лабинцев. – Режим доступу: https://habr.com/ru/company/ods/blog/328372/
3. Narkhede, S. Understanding Confusion Matrix [Електронний ресурс] / S. Narkhede. – Режим доступу: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
4. Афанасьев, С. Gini & ROC & Precision-Recall: проблемы метрик в банковском моделировании [Електронний ресурс] / С. Афанасьев, А. Смирнова. – Режим доступу: http://futurebanking.ru/post/3761
5. Encyclopedia of Machine Learning / C. Sammut and G.I. Webb, Eds. – New York: Springer, 2011. – 892 p.
6. Statlog (Heart) Data Set [Електронний ресурс]. – Режим доступу: http://archive.ics.uci.edu/ml/datasets/statlog+(heart).
7. Van der Maaten, L.J.P.. Visualizing High-Dimensional Data Using t-SNE / L.J.P. van der Maaten, G.E. Hinton // Journal of Machine Learning Research. – 2008. – № 9. – P. 2579 – 2605.
8. Supervised learning [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/supervised_learning.html
9. Піскун, О.В. Застосування методів машинного навчання для побудови моделі рішення задачі класифікації / О.В. Піскун // Вісник Черкаського університету. Серія: Прикладна математика. Інформатика. – 2019. - №1. – С. 41-52.

References:

1. Delgado, M.F., Cernadas. E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research,* 15, 3133-3181.
2. Labintsev, E. (2017). Metriki v zadachah mashinnogo obucheniya [Machine Learning Metrics]. Retrieved from https://habr.com/ru/company/ods/blog/328372/ [in Russian].
3. Narkhede, S. (2018). Understanding Confusion Matrix. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

4. Afanasev, S., Smirnova, A. (2019). Gini & ROC & Precision-Recall: problemy metrik v bankovskom modelirovanii [Gini & ROC & Precision-Recall: problems of metrics in banking modeling]. Retrieved from http://futurebanking.ru/post/3761 [in Russian].
5. Sammut, C. & Webb, G.I. (Eds.). (2011). Encyclopedia of Machine Learning. New York: Springer.
6. Statlog (Heart) Data Set. (n.d.). Retrieved from http://archive.ics.uci.edu/ml/datasets/ statlog+(heart).
7. Van der Maaten, L.J.P., & Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research,* 9(Nov), 2579–2605.
8. Supervised learning. (2019). Retrieved from https://scikit-learn.org/stable/supervised_learning.html
9. Piskun, O.V. (2019). Zastosuvannya metodiv mashinnogo navchannya dlya pobudovi modeli rishennya zadachi klasifikaciyi [Application of machine learning methods to build a model for solving the classification problem]. Visnik Cherkaskogo universitetu. Seriya: Prikladna matematika. Informatika – Cherkasy University Bulletin: Applied Mathematics. Informatics, 1, 41–52 [in Ukrainian].

PISKUN Oleksandr,
Candidate of Technical Sciences, Associate Professor, Chair of Department of Applied Mathematics and Informatics, Bohdan Khmelnytsky National University of Cherkasy

ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR THE BINARY CLASSIFICATION PROBLEM

Summary. Introduction. The analysis of existing common methods of classification applied to the problem of the heart disease diagnostics was provided in the present paper. Various quality metrics of binary classification models that can be used in deciding on the optimality of the developed model are considered. Modelling was conducted with and without parameter optimization. Parameter tuning was performed using validation curves, followed by grid search with cross-validation of each parameter combination. Decision Tree Classifier, Gradient Boosting Classifier and Gaussian Naive Bayes showed the best results for our particular problem.

Binary classification is one of the most common problems of applied statistics and machine learning, which is present in many applied fields - in medicine, biology, meteorology, analysis of mail messages, classification of texts, images, etc.

Assessment of the quality of classification models is an important aspect in all areas to which machine learning models are applied. Accuracy Score answers the question of how well a classifier separates the classes we are interested in over some sample.

Purpose. The purpose of the present paper is to find the optimal machine learning algorithm to solve the problem of classification for determining the presence or absence of heart disease on the basis of real data.

Results. A set of commonly used classifiers is investigated on real data of heart disease. Models accuracy scores (with 10 windows cross-validation) with default and optimized parameters were obtained. The accuracy metric was generally used to evaluate the performance of the model, and recall score - for our particular problem.

Conclusion. The analysis of common methods of classification applied to the problem of the heart disease diagnostics was carried out in the present work. Different quality metrics of binary classification models effectiveness that can be used in deciding on the optimality of the model were considered. Decision Tree Classifier, Gradient Boosting Classifier and Gaussian Naive Bayes showed the best results for our particular problem, but they are still not effective enough to be applied in the medical sector. Since we have a balanced dataset with no outliers, it is most likely that the sample is not large enough for the optimal model training.

Future studies will provide a detail analysis of the selected models, in particular, finding reasons for the volatility of recall metric values to reduce it and, consequently, improve the stability of model accuracy. Based on these three methods, an assembly classification of heart disease recognition will be built, which might improve the accuracy of the model.

Keywords: machine learning, accuracy metrics, binary classification, classification algorithms.

Одержано редакцією 22.05.2019 р.
Прийнято до публікації 09.10.2019 р.