УДК 004.93

# ХАРАКТЕРИСТИКА ДЛЯ ВИБОРУ МОДЕЛЕЙ У АНСАМБЛІ КЛАСИФІКАТОРІВ

*О.В. Бармак, Ю.В. Крак, Е.А. Манзюк*

Проведено аналіз досліджень та підходів практичного застосування ансамблів та визначено характерні фактори впливу на комбінацію моделей. Фактори несуть визначальний характер та притаманні комбінаціям застосувань. Обгрунтовано необхідність використання ознак моделей, які характерні тільки для ансамблів. Встановлені характерні особливості ансамблів та визначено необхідність розробки специфічних характеристик застосувань в розрізі комбінації рішень з визначення їх характерних ознак. Ці ознаки, а саме точність та відмінність здійснюють визначальний вплив на вибір та застосовність рішень в ансамблях і дозволяють вибрати найбільш дієву комбінацію. Запропоновано до використання таку характеристику рішення в ансамблі як відмінність певного рангу за параметром точності. Ця характеристики моделей дозволяє здійснювати їх вибір та характеризує модель в ансамблі. Вона застосовна тільки у випадку комбінації моделей. Вказує на відмінність однієї моделі від іншої та враховує точність моделі. Застосовується для моделей різної природи. Дозволяє визначити глибину відмінності моделей та дозволяє поєднуватись і з іншими відомими характеристиками класифікаторів. Особливість її полягає в тому, що вона дозволяє дати оцінку використання рішень в ансамблі та здійснювати вибір рішень.
Ключові слова: ансамблі, розмір ансамблів, відмінність, точність.

Проведен анализ исследований и подходов практического применения ансамблей и определены характерные факторы влияния на комбинацию моделей. Факторы несут определяющий характер и присущи комбинациям приложений. Обоснована необходимость использования признаков моделей, которые характерны только для ансамблей. Установлены характерные особенности ансамблей и определена необходимость разработки специфических характеристик приложений в разрезе комбинации решений по определению их характерных признаков. Эти признаки, а именно точность и отличие осуществляют определяющее влияние на выбор и применимость решений в ансамблях и позволяют выбрать наиболее действенную комбинацию. Предложено к использованию такую характеристику решение в ансамбле как различие определенного ранга по параметру точности. Эта характеристики моделей позволяет осуществлять их выбор и характеризует модель в ансамбле. Она применима только в случае комбинации моделей. Указывает на различие одной модели от другой и учитывает точность модели. Применяется для определения моделей различной природы. Позволяет определить глубину различия моделей и позволяет сочетаться и с другими известными характеристиками классификаторов. Особенность ее заключается в том, что она позволяет дать оценку использования решений в ансамбле и осуществлять выбор решений.
Ключевые слова: ансамбли, размер ансамблей, различие, точность.

The analysis of researches and approaches of practical application of ensembles is carried out and the characteristic factors of influence on a combination of models are determined. Factors are determinative and inherent in combinations of models. The necessity of using patterns of models that are characteristic only for ensembles is substantiated. The characteristic features of ensembles are established and the necessity of developing specific characteristics of models in terms of a combination of solutions by definition of their characteristic features is determined. These attributes, namely precision and distinction have a decisive influence on the choice and applicability of solutions in ensembles and allow to choose the most effective combination. We propose to use such a characteristic of the models in the ensemble as distinction of a certain rank by precision parameter. These characteristics of the models allow them to be chosen and characterize the model in the ensemble. It applies only in the case of a combination of models and indicates the distinction between one model and another and takes into account the precision of the model. This approach allows you to define models of different nature, to determine the depth of model distinction and allows it to be combined with other well-known characteristics of classifiers. Its peculiarity consists in the fact that it allows to evaluate the use of solutions in the ensemble and to carry out the selection of models.
Key words: ensembles, ensemble size, distinction, precision.



## Вступ

В електронному вигляді знаходиться широкий спектр текстової інформації, як то, публікації, періодичні видання, журнали, книги а також ті види, що мають тільки електронне подання, – пости, щоденники, повідомлення, відгуки тощо. Люди все частіше в пошуках необхідної їм інформації з текстових ресурсів використовують електронне подання інформації. З метою задоволення інформаційних потреб виникає проблема керованості зростаючим об'ємом текстових джерел. Ряд підходів щодо досягнення ефективного розділення інформації базуються на різних етапах підготовки даних, методах відбору ключових ознак, методах інтелектуального аналізу даних, методиках оцінки результатів для певного набору даних тощо.

Класифікація тексту є завданням встановлення відповідності належності змістовності тексту до визначених категорій. Формально це можна подати у вигляді існування документу $x_i$ з набору документів $X$ та набору категорій $\{y_1, y_2, y_3, ..., y_n\}$ і встановлення відповідності категорії документу $x_i \in y_j$, при зазначених певних критеріях відповідності. Це встановлює строгу відповідність категорії документу та відповідає бінарній класифікації.

Документи у співвідношенні з критеріями відповідності можуть бути позначені міткою приналежності одного класу або більш ніж одного. Якщо документ може бути віднесений до категорій більш ніж одного класу виникає проблема мультикласовості, що може розглядатися як певний вид згортки з подальшим

вирішенням задачі бінарної класифікації. Завдання класифікації зводиться до таких послідовних етапів як подання документу, вибір функції або ознак, застосування підходів інтелектуального аналізу даних та оцінка ефективності.

Автоматична класифікація текстів є важливим напрямком та будується за методами машинного навчання, що автоматично конструюють класифікатор виходячи з характеристик категорій з набору попередньо класифікованих документів. Ця обставина відіграє важливу роль в отримані та узагальнені інформації. Якщо використовувати весь загал електронного подання текстової інформації, то, як правило, вона носить неоднорідний характер. Оскільки інформація збирається здебільшого на просторах Інтернету з друкованих новин, статей, електронних дошок оголошень, конференцій, форумів, реклами оглядів фільмів тощо, та має різні формати, різні мови, різні стилі написання. Таке різноманіття у подані зумовлює всю важливість та необхідність систем автоматичної класифікації. В зв'язку з цим, питанню класифікації присвячено велику кількість досліджень та розроблено широкий спектр підходів, що дають досить часто порівнюванні результати на одній множинні даних.

Однак вибір рішень та алгоритмів інтелектуального аналізу для вирішення поставлених завдань є складним та містить ряд проблем [1]. Однією з них є широке різноманіття підходів та рішень, які напрацьовані та можуть бути застосовні. В зв'язку з цим необхідно піддавати данні певним перетворенням, тестувати алгоритми та підходи, підбирати параметри з метою пошуку реалізацій, які дають прийнятні результати. Як відомо, поєднанням декількох алгоритмічних рішень дає більш кращі результати і носить назву ансамбль алгоритмів. Дослідження із використання проводяться і сьогодні [2], хоча даний підхід відомий досить давно. Так як не існує ідеального алгоритму, використовують критерії оцінювання його роботи, як приклад, повнота, точність та інші. Оскільки будь-який алгоритм має свої сильні та слабкі стороні, їх поєднання в ансамблі в значній мірі ускладнює дослідження та практичне застосування. Питання вибору критеріїв компонування алгоритмів досі не вирішене. У випадку ансамблю необхідні критерії, які дозволяють охарактеризувати алгоритм не як окремо працюючу модель, а в системі взаємодії з іншими рішеннями. Тобто необхідно розробити інформативні критерії оцінювання моделі в розрізі її застосування в ансамблі. При цьому ці характеристики не повинні інформативно дублювати відомі та застосовні, а їх доповнювати та використовуватись у взаємодії з ними.

## Постановка задачі

Завданням є розробка інформативних характеристик моделей при комбінації рішень в ансамблі, які дозволяють визначити параметри взаємодії і унікальності застосувань та критерії їх вибору для використання у ансамблі. Ці параметри повинні носити виняткову інформативність та доповнювати і поєднуватись у застосуванні з відомими характеристиками рішень.

## Методи оцінки класифікаторів

Є декілька підходів щодо подання документів. Досить часто використовується підхід Bag-of-words (набір слів), де тексти подаються у вигляді наборів слів без врахування їх розташування та зв'язків. Багато з цих слів не носять семантичного навантаження і загальну кількість слів можна зменшити, зменшуючи таким чином простір ознак. Одним з найбільш відомих способів зменшення списку слів є вилучення слів, що не впливають на класифікацію документів, так званих «стоп-слів». Інший підхід полягає у зменшенні кількості слів використовуючи базові словоформи (основи) слова (stemming). Такі підходи хоча і зменшують кількість слів однак загальна кількість залишається досить значною. Вибір множини ознак для текстового документа застосовується для кожного слова. Підрахунок ознак для кожного слова або індивідуальна особливість слова може бути подана різними підходами (частотою документа, частотою терміну, $\chi^2$ статистикою, силою терміну тощо). За цими підходами можна проводити ранжування незалежно одне від одного і вибрати характеристики, що дають найбільш значимі параметри. Інший підхід полягає у застосуванні функціонального перетворення та видалені ознак. Застосування методу головних компонент дозволяє зменшити розмірність даних. Тобто метою є зменшення розмірності даних та розбіжності шляхом видалення частини нерелевантної текстової інформації в розрізі класифікації на множині відомих класифікаторів втративши при цьому найменшу кількість інформації.

Векторна модель семантики може бути подана у вигляді матриць різних типів. Матричний підхід дозволяє сформувати такі матриці як термін-документ, де оцінювання релевантності проводиться у вигляді набору слів; слово-контекст, використовується для знаходження подібності окремих документів або частин документів; пара-модель, згідно якої пари слів в подібних моделях мають близьку семантичну залежність та близькість векторів-рядків матриці.

Методи класифікації можуть базуватися на статистичних підходах та методах машинного навчання з використанням індуктивного та дедуктивного підходів. Вони досить суттєво різняться архітектурою та прийнятими підходами.

Оцінка якості класифікаторів має експериментальний характер, що базується на неформалізованості та суб'єктивності якості результатів. Базовими характеристиками якості є рівність помилок першого та другого

роду. До помилок першого роду відносять хибний пропуск, тобто документ, який належить класу не було класифікатором до нього віднесено. Помилки другого роду – хибна класифікація, позначається у випадках коли документ було віднесено до нерелевантного класу. Найбільш часто з метою оцінки ефективності використовують статистичні параметри точність (precision) та повнота (recall), що відносяться до задачі бінарної класифікації з множини $\{0,1\}$.

З використанням матриці помилок (табл. 1) визначаються параметри:

$$\begin{aligned} Точність &= \frac{TP}{TP+FP}; \\ Повнота &= \frac{TP}{TP+FN}. \end{aligned} \qquad (1)$$

Характерним є те, що точність та повнота не залежать від співвідношення розмірів класів.

Таблиця. Матриця помилок

| | | Оцінка експерта (дійсна) | |
|---|---|---|---|
| | | Позитивна | Негативна |
| Оцінка класифікатора | Позитивна | True Positive (TP) | False Positive (FP) |
| | Негативна | False Negative (FN) | True Negative (TN) |

В текстовій класифікації зазвичай присутні шумові елементи. Особливість їхнього впливу полягає в тому, що вони деформують інформаційне ядро класу. Шумовою ознакою називають таку ознаку, включення якої в документ в середньому підвищує помилку класифікації. Шумовий документ у разі включення його в навчальну множину підвищує помилку класифікації. Базовий розподіл розділює простір документів на підмножини з однорідними ознаками класів, і якщо документ потрапив в підмножину в якій він не відповідає ознакам класу, тоді він є шумовим. Наявність таких документів є проблемою завдяки якій сильно ускладняється завдання з навчання класифікатора. При зосередженні уваги на шумових документах під час вибору розділювальної границі класифікатор стає неточним для нових даних. Досить часто важко визначити які документи можуть бути шумовими.

При розділенні класів лінійним класифікатором та знаходженні лінійного розділювача існує проблема його знаходження. Причина полягає в тому, що при наявності гіперплощини, що ідеально розділює класи існує безмежність лінійних розділювачів. Якщо класи лінійно розділювальні, важливим є вибір критерію для знаходження розділювальної гіперплощини серед множини таких гіперплощин, які ідеально розділюють дані для навчання. Вибір необхідно здійснити таким чином, щоб ця гіперплощина також досить добре розділювала нові дані.

У випадку неможливості вирішення завдання лінійним підходом та визначення границі за допомогою гіперплощини, використовують нелінійні класифікатори. В таких випадках нелінійні класифікатори точніше визначають приналежність класам за лінійні. Якщо ж завдання можна вирішити лінійним підходом, краще вирішити лінійною класифікацією. Однак використання нелінійного класифікатори може також говорити і про те, що набір ознак підмножин розділювальних класів є неякісним з точки зору лінійної розділювальності та необхідно провести ревізію ознак.

Зважаючи на те що процес класифікації є поетапним, важливим є вибір підходу до кожного етапу та встановлення параметрів і критеріїв якості кожного з етапів. Вибір алгоритму інтелектуального аналізу даних, методу вибору функцій, міри оцінки повинен проводитись з можливістю корегування та повернення на попередні етапи. На вибір класифікатора суттєво впливає обсяг даних і метод навчання. Будь-який класифікатор має свої переваги та недоліки, що проявляються і в залежності від обсягу даних. Проте при великих обсягах вибір класифікатора незначною мірою впливає на результати класифікації і вибір найкращого класифікатора.

Тому для перевірки та порівняння використовують однакові набори даних, що поділені на навчальну та тестуючи частину. Досить широко для досліджень використовують корпус текстів Reuters-21578, RCV1 (Reuters Corpus Volume 1), TREC-AP.

## Комбінації моделей в ансамблі

Природнім є підхід, який полягає в тому що при недостатній здатності класифікатора узагальнити ознаки класифікації можна змінити параметри попередніх етапів процесу або поєднати декілька

класифікаторів. Однак на нашу думку важливим є визначення множини даних, які "важко" піддаються класифікації, тобто не класифікуються переважною більшістю класифікаторів.

Метод бустінг полягає у тому, що використовують послідовно ряд класифікаторів з впливом вихідного класифікатора на наступні в ланцюгу. Після опрацювання вхідної навчальної множини класифікатором здійснюється перевірка і перерахунок коефіцієнтів документів. Коефіцієнт зменшується за умови вірної класифікації документа та збільшується при хибній класифікації. У бустінгу використовують зважену лінійну комбінацію голосів ансамблю. Метод базується на жадібному алгоритмі побудови композиції алгоритмів. Такий підхід є популярним поряд з методами штучного інтелекту та методом опорних векторів.

Незалежну та паралельну класифікацію використовує баггінг. Навчання здійснюється на множинах отриманих з вихідної множини шляхом випадкової заміни та повторювання документів в множині. Простою більшістю голосів класифікаторів визначають результат класифікації. Комбінації незалежних класифікаторів є більш ефективними для використання в ансамблі класифікаторів.

Баггінг (bagging) використовує підмножини із заміщенням для навчання на базі навчальної вибірки. Елементи підмножин можуть перетинатись та дублюватись і результат є більш точнішим ніж при використанні тільки однієї базової навчальної вибірки. Дозволяє поєднати оцінку передбачення класифікаторів навчених з використанням випадкових підмножин базової навчальної вибірки.

В загальному випадку розглядають задачу побудови функції передбачення $\hat{f}(x): \mathrm{X} \to \mathrm{Y}$. Для навчальної вибірки виникає необхідність встановлення передбачуваної відповідності простору об'єктів $\mathrm{X} = \{x_1, x_2, ..., x_N\}$ простору відповідей цільової змінної $\mathrm{Y} = \{y_1, y_2, ..., y_N\}$, де $N$ загальна кількість пар відповідностей, що використовуються для навчання. На базі навчальної вибірки $\mathrm{Z} = \{z_1, z_2, ..., z_N\}$, $z_i = (x_i, y_i)$, $i = 1, 2, ..., N$ отримуємо оцінку передбаченні $\hat{f}(x)$ на вхідних $x$. Агрегація бустінга для отримання середньої оцінки над колекцією результатів баггінга дозволяє зменшити дисперсію. Кожна оцінка передбачення визначається як усереднене значення на базі функціональних оцінок навчених на випадкових підмножинах вибірки $\mathrm{Z}^{*c} \subset \mathrm{Z}$ $c = 1, 2, ..., C$ , де $C$ загальна кількість вибірок, яка відповідає кількості функціональних оцінок.

Загальна оцінка баггінга

$$\hat{f}_b(x) = \frac{1}{C} \sum_{c=1}^{C} \hat{f}^{*c}(x). \tag{2}$$

Позначимо $\hat{P}$ емпіричний розподіл вибору підмножин еквівалентної ймовірності $1/N$ для кожної з пар відповідності $(x_i, y_i)$. Реальна оцінка визначається $\mathrm{E}_{\hat{P}} \hat{f}^*(x)$ при $Z^* = \{(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_N^*, y_N^*)\}$ і для кожної пари $(x_i^*, y_i^*) \sim \hat{P}$. Таким чином реальна оцінка визначається $\hat{f}_b(x) \to \hat{f}(x)$ при $C \to \infty$.

Стекінг (stacking) використовується в ансамблі із застосуванням мета-алгоритму. Підмножини, які не перетинаються формуються на базі навчальної вибірки. Вихід одного класифікатора використовується як навчальна множина для іншого класифікатора апроксимуючи ту ж цільову функцію формуючи таким чином мета-ознаку.

Класифікатори вибираються як одного типу так і можуть бути різної природи. Можемо мати набір класифікаторів, які кореспондуються з функціональною оцінкою, $\mathrm{K}_k$ де $k = 1, 2, ..., \mathrm{K}$ для навчальної вибірки $\mathrm{Z}$. Апостеріорний розподіл кількісної оцінки $\zeta$ функції передбачення $f(x)$ на фіксованій кількості ознак змінної $x$ при розподілі $\Pr(\mathrm{K}_k \mid \mathrm{Z})$ визначається залежністю

$$\Pr(\zeta \mid \mathrm{Z}) = \sum_{k=1}^{\mathrm{K}} \Pr(\zeta \mid K_k, \mathrm{Z}) \Pr(\mathrm{K}_k \mid \mathrm{Z}). \tag{3}$$

Бустінг на дереві рішень є ефективним методом класифікації і покращує якість класифікації з ростом композиції та зменшення помилок. Класифікатори композиції можна розглядати як функції рішень результати яких при їх кількісному збільшені лінеарізують бінарну роздільність кінцевого результату.

Навчання класифікатора здійснюється на базі підмножини навчальної вибірки. Підмножина замінюється фіксованою розмірністю та замінюється $K$ разів. Класифікатор перенавчається для кожного з набору підмножин та перевіряється на відповідність на кожному з $K$ повторень. Оціночна варіативність обчислень $S(\mathrm{Z})$ на навчальній вибірці оцінюється таким чином

$$V[S(\mathrm{Z})] = \frac{1}{K-1}\sum_{k=1}^{K}\left(S\left(Z^{*k}\right) - \bar{S}^{*}\right)^2 \ , \tag{4}$$

де $\bar{S}^{*} = \sum_{c=1}^{K} S\left(Z^{*c}\right) \Big/ K$.

На базі цього визначається похибка оцінки передбачення, яка завдяки крос-валідації зменшує свою величину.

Стекінг дозволяє визначити оцінку передбачення змінної $x$ використовуючи класифікатор $c$ на наборі даних з видаленням $i$ -го елемента та ґрунтується на функціональній оцінці на базі крос-валідації.

Визначаються вагові коефіцієнти

$$\hat{\omega} = \arg\min_{\omega} \sum_{i=1}^{N}\left[ y_i - \sum_{c=1}^{C} \omega_c \hat{f}_c^{-i}(x_i) \right]^2 . \tag{5}$$

Результуюча оцінка

$$\hat{f}_s = \sum_{c=1}^{C} \hat{\omega}_c \hat{f}_c(x). \tag{6}$$

Така оцінка близька до поелементної крос-валідації що дозволяє вибрати на базі вагових коефіцієнтів найкращий класифікатор оцінки $\hat{f}$ з найменшою помилкою. Як при застосуванні класифікаторів при змінні вхідних параметрів так і при застосуванні комплексів класифікаторів оцінюють результати класифікації через співвідношення до загального простору, що не дає можливість визначити інформаційне ядро та встановлювати відмінності параметрів. Важливим є застосування показника безпосереднього порівнювання іменованих результатів класифікації з метою встановлення подібності порівнювальних параметрів.

Для порівняння роботи класифікаторів використовуються параметри порівняння. Оцінка ефективності класифікатора здійснюється на базі мікроточностості $p$, мікроповноти $r$ та $F$ міри на множині класів $L$ та множині документів $D$. Параметри є модифікованими представлення підходів (1)

Мікро оцінка

$$p = \frac{\sum_{l=1}^{|L|}\left|\hat{D}_l \cap D_l\right|}{\sum_{l=1}^{|L|}\left|\hat{D}_l\right|}; \quad r = \frac{\sum_{l=1}^{|L|}\left|\hat{D}_l \cap D_l\right|}{\sum_{l=1}^{|L|}\left|D_l\right|}. \tag{7}$$

Макро оцінка

$$p = \frac{1}{|L|}\sum_{l=1}^{|L|}\frac{\left|\hat{D}_l \cap D_l\right|}{\left|\hat{D}_l\right|}; \quad r = \frac{1}{|L|}\sum_{l=1}^{|L|}\frac{\left|\hat{D}_l \cap D_l\right|}{\left|D_l\right|}. \tag{8}$$

Для спільної оцінки повноти та точності використовується $F$ параметр

$$F = \frac{2pr}{p+r}. \tag{9}$$

Як можна помітити завдання порівняння класифікаторів розглядається з кількісної точки зору. Класифікація здійснюється за множиною ознак з якою працює класифікатор та співвідносить документ з класом на базі ознак.

## Характеристика моделей в ансамблі

Використання декількох підходів для рішення задачі машинного навчання є досить популярним зважаючи на те, що такий підхід досить очевидний і лежить на поверхні. Крім того сфера машинного навчання є об'ємною за кількістю підходів та швидко розвивається пропонуючи все нові модифікації відомих підходів, таким чином все більше розширюючи поле можливих застосувань. Зважаючи на широке різноманіття рішень завдання вибору комбінації підходів як самих рішень в ансамблі так і способів рішень моделей в самих ансамблях, які теж є різноманітними за своїми підходами, виникає проблема вибору застосувань. Власне кажучи існує проблема вибору рішень для вирішення практичних задач [1]. Одно із головних проблем полягає в тому, що практично неможливо в простий спосіб використати відомі підходи для вирішення конкретного завдання. Немає чітких підходів за якими можна було б встановить в яких випадках одне рішення працює краще від іншого. І рекомендації зводяться до перебору великої кількості рішень із комбінацією параметрів. Суттєво на результат роботи мають вплив і данні до яких застосовуються рішення. Крім того існує проблема практичних реалізації підходів та впливу параметрів на результати роботи. Звісно такі ж проблеми притаманні і ансамблям, та в достатній мірі вони поглиблюються породжуючи нові.

Однак поєднання рішення та застосування їх в ансамблях знаходить своє застосування з практичними результатами і сьогодні [2]. Тому незважаючи на складність використання ансамблів є актуальним та потребує вирішення проблеми вибору рішення для їх комбінації. Використання комбінації залежить від рішень, які використовуються і не полягають в простому поєднанні. Для різних рішень необхідні різні ознакові простори і потребують попередньої обробки даних.

Результат по роботі ансамблю визначається теж по різному. Так в задачі регресії як середнє значення а в задачі класифікації як голосування за більшістю, і результат часто за якістю переважає рішення, які входять в ансамбль. Досить часто говорять про використання рішень різної природи [3], тобто використання рішень з різними техніками, поєднання яких дає більш кращий результат. Як приклад, поєднання лінійних, метричних технік, на базі дерев, тощо. Проте формалізація цього питання досить складна та мало досліджена [4, 5]. Відомі роботи [4], які пропонують сімейство евристик із визначення відмінності підходів, де автори називають diversity. Це можна перекласти як різновидність або різноманітність. Щоб побудувати хороший ансамбль, необхідно не тільки побудувати хороші базові класифікатори, а також класифікатори повинні бути відмінні, що значить для того самого прикладу база класифікаторів повертає різні виходи і їхні помилки повинні бути в різних прикладах [4]. Методи ансамблю відрізняються таким чином, як вони визначають відмінність між базовими класифікаторами. Загалом відмінність між класифікаторами повинна бути якомога кращою, тобто необхідно максимізувати функцію відмінності. Найбільш критичним є поняття відмінності при використанні стекінга. В цьому випадку рішення приймаються метаалгоритмом, на базі одно з видів голосування (voting). Вибір за голосуванням приймається як функція експертних думок, в даному випадку класифікаторів.

Важливим є визначення характеристик класифікатора, тобто параметри за якими його можна було б оцінити. Відомо ряд параметрів, таких як повнота, точність (1) та інші (7, 8, 9), за яким можна дати оцінку класифікатору. Найбільшу популярність знайшов ROC - крива помилок та кількісна оцінка на базі ROC, AUC - площа, обмежена ROC-кривою та віссю FPR (хибною позитивною класифікацією).

Однак нам необхідний показник який би оцінював не сам підхід, а підхід в контексті ансамблю, тобто давав ознаку за якою можна оцінити доцільність його використання і мав нескладне представлення та був носієм інформативності в напрямку практичної застосовності. Водночас володів властивістю агрегативності та узагальненості з позиції поєднання з іншими застосовними показниками, як наприклад AUC. При цьому необхідно забезпечити відсутністю дубльованості інформативності у поєднанні показників.

Так як при голосуванні за експертними оцінками визначальним є рішення експерта, необхідний показник, який би характеризував експерта в контексті ансамблю і тільки в цьому контексті. В цьому контексті і розглянемо якості експерта. Експерт в переважній більшості повинен давати правильні рішення, тобто вірно класифікувати базові дані. Відповідно до цього необхідно навчати на розмічених даних, для визначення правильності класифікації. Однак особлива цінність експерта в контексті ансамблю полягає не тільки в правильності рішення, а також у відмінності його думки від інших. Тобто думка експерта повинна бути правильною та відмінною. В площині класифікації, класифікатори повинні правильно класифікувати розмічені дані, і с особливо важливо правильно класифікувати дані, які некласифіковані (тобто неправильно класифіковані) іншими класифікаторами. Таким чином необхідно мінімізувати функцію помилок та максимізувати функцію відмінності. І ці параметри необхідно покласти в основу показника експерта тобто класифікатора.

Визначимо пріоритетність параметрів відмінності та правильності. З точки зору використання рішень в ансамблі важливе значення має особливість думки експерта. Власне кажучи, ансамблі і використовують з метою компенсації недоліків окремих думок, і в кінцевому випадку неправильність думки окремого експерта компенсується існуванням ансамблю. Водночас відмінність окремої думки відіграє важливу роль саме у формування ансамблю та формування метаознак. У випадку вибору метаознак добре натренованих моделей виникають проблеми з сильною кореляцією думок. В цьому випадку базові рішення не сильно оптимізують або застосовують інші евристичні підходи. Іншими словами використовувати двох експертів, думки яких завжди майже співпадають не має особливого сенсу. Достатньо використати одного а не серію дуже подібних

експертів за своїми думками. Тому переважаюче значення в цьому випадку має відмінність правильної думки експерта, а неправильність думки компенсується існуванням самого ансамблю.

Крім того показник повинен бути застосовним для визначення характеристик кількості рішення в ансамблі. Розмір ансамблю відіграє досить часто критичну роль [6], так як поєднання різних підходів є ресурсоємним. І завдання полягає у визначенні розміру ансамблю, який достатній для вирішення конкретного завдання. Дослідження вказують про можливість формування оптимального розміру за критерієм точності кількості компонентів ансамблю для вирішення завдання. Теоретичні дослідження показують, що використання такої ж кількості незалежних рішень як міток класів дає найвищу точність [6]. При збільшені розміру ансамблю віддача системи в розрізі точності знижується. Однак і в цьому застосуванні необхідно використовувати якомога відмінні рішення.

Таким чином використання відмінності є важливим у композиції рішень. Водночас необхідно використовувати рішення, які дають максимально високу точність, що і необхідно поєднати в одному показнику. Поєднання точності та відмінності рішення є контроверсивним. Це пояснюється там що функція максимізації точності призводить до зниження відмінності на множині розмічених даних. І звичайно навпаки, при значній відмінності, точність думок буде знижуватись. Саме це поєднання в одному показникові є особливо важливим за приорітетністю відмінності.

Відмінність експерта полягає в тому наскільки його думка відмінна від дутки інших. Застосовно до класифікаторів визначається за результатами відмінності правильних класифікацій на просторі розмічених даних $U$ (рисунок).

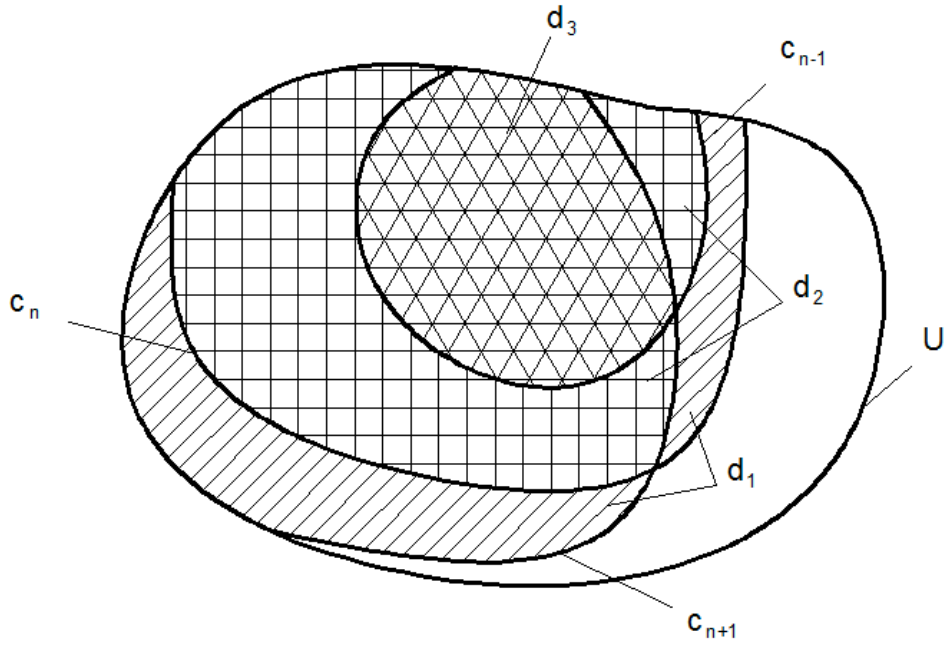


Рисунок. Множини вірних класифікацій

Якщо результати класифікатора є унікальні, тобто є вірними та не співпадають з жодним іншим зони $d_1$ множин правильних результатів класифікаторів $c_n$ та $c_{n+1}$, відмінність є суттєвою і необхідно ввести поняття рангу відмінності $r$, а відмінність позначати з рангом $d_{r=1}$. Зміст відмінності з рангом, наприклад один, полягає в тому, що наявна множина вірних класифікацій класифікатора $c$ і жоден елемент цієї множини не належить будь-якій множині вірних відповідей іншого класифікатора. Відповідно при зменшені вимог відмінності до класифікатора можна послабити ранг відмінності і допустити щоб класифікатор дав правильний результат, який співпадає ще з одним в тільки одним із множини класифікаторі $C$. Тобто в цьому випадку ранг $r = 2$ та отримаємо зони позначені $d_2$ (див. рисунок). Таким чином послаблення вимог до класифікатора за критерієм відмінності можна здійснювати з встановленням рангу відмінності. А визначати відмінність класифікатора відповідно до рангу як об'єднання множин класифікатора до відповідного рангу з урахуванням більш сильніших рангів класифікатора $\bigcup^{r} d_{cr}$.

Для визначення параметра експерта проіндексуємо універсальний простір розмічених даних $U$ та застосуємо до нього набір класифікаторів $\{c\}$ за правильними результатами роботи класифікованих даних

$X \times Y$. Відмінність кожного класифікатора визначимо як відношення даних правильно класифікованим цим класифікатором до даних, які правильно класифіковані також і іншими класифікаторами відповідно до рівня рангу. Співвідношення визначаємо в межах множини класифікатора. Результати обрахунку можуть змінюватися в широких межах.

**Input:** індексація розмічених даних $U$
**Set:** $r$
**forall** $c \in C$ **do**

$$t \leftarrow \frac{\left|\left\{ x^y \mid x^y \in \bigcup^{r} \bigcap^{C} C_c^r \setminus C_c \right\}\right|}{\left|\left\{ x^y \mid x^y \in \bigcup^{r} \bigcap^{C} C_c^r \right\}\right|}; \quad (10)$$

$$p \leftarrow \frac{\left|\{C_c \setminus U\}\right|}{\left|\{U\}\right|}; \quad (11)$$

$$d_{c(r)} \leftarrow t/p\,. \quad (12)$$

**end**
**sort** $\{d_{c(r)}\}$
**return** $d_{c(r)}$

Параметр точність визначимо як відношення множини неправильно класифікованих даних до універсальної множини даних (11). Цей параметр має відносний характер з максимальним значення $p < 1$. Співвідношення (12) призначене для підсилення відмінності у відповідності до точності, та дозволяє поєднати в одному параметрі ці два значимих фактора. Таким чином визначається така характеристика рішенням в ансамблі як відмінність за параметром точності. Відмінність визначається за встановленим рангом $r$. Вона дозволяє врахувати як ранг відмінності рішення так і точність. Саме подвійна назва в повній мірі визначає призначення характеристики. Таким чином отримаємо таку характеристику класифікатора як ***відмінність рангу*** $r$ ***за параметром точності***. Ця характеристика призначена для використання в ансамблі та може бути використана як в безпосередньому вигляді (10) – відмінність, так і за параметром точності (12). Також можна використовувати будь-який інший параметр, який застосовний для характеристик рішень та змінюється в межах (0, 1). Найбільш часто він носить відносний характер. Як приклад, можна привести використання відмінності рангом $r$ за параметром ROC.

## Висновки

Аналіз досліджень та підходів практичного застосування ансамблів дозволив визначити характерні фактори впливу на комбінацію рішень, які несуть визначальний характер та притаманні комбінаціям застосувань. Встановлені характерні особливості ансамблів та визначено необхідність розробки специфічних характеристик застосувань в розрізі комбінації рішень з визначення їх характерних ознак. Ці ознаки, а саме точність та відмінність здійснюють визначальний вплив на вибір та застосовність рішень в ансамблях і дозволяють вибрати найбільш дієву комбінацію. Запропоновано до використання таку характеристику рішення в ансамблі як ***відмінність певного рангу за параметром точності***. Ця характеристика доповнює відомі та застосовні і використовується поряд з іншими та в об'єднанні з ними. Особливість її полягає в тому, що вона дозволяє дати оцінку використання рішень в ансамблі та здійснювати вибір рішень.

## Література


1. Brownlee J. (2014) A Data-Driven Approach to Choosing Machine Learning Algorithms. [Online] September 29th 2014. Available from: Machinelearningmastery.com https://machinelearningmastery.com/a-data-driven-approach-to-machine-learning/. [Accessed: 30 January 2018].
2. Ren S., Liao B., Zhu W., Li K. (2018) Knowledge-maximized ensemble algorithm for different types of concept drift. Information Sciences. 430–431 (3). P. 261–281.
3. Hsu C-H., Shi X., Salapura V. (2014) 11th IFIP International Conference on Network and Parallel Computing (NPC), 18-20 September 2014. Ilan, Taiwan. Springer. LNCS-8707. P. 333–346, 2014. Network and Parallel Computing.
4. Diez-Pastor J.F., Rodríguez J.J., García-Osorio C. & Kuncheva L.I. (2015) Diversity techniques improve the performance of the best imbalance learning ensembles. Information Sciences. 325. P. 98–117.
5. Löfström T. (2015) On Effectively Creating Ensembles of Classifiers: Studies on Creation Strategies, Diversity and Predicting with Confidence. A Thesis Submitted in partial fulfilment of the Requirements of Stockholm University for the Degree of Doctor of Philosophy. Stoke-on-Trent: Stockholm University.

6. Hamed R.B., Fazli C. (2017). Less Is More: A Comprehensive Framework for the Number of Components of Ensemble Classifiers. IEEE Transactions on Neural Networks and Learning Systems. [Online] 14(8), September 2017 USA: IEEE. P. 1–7. *Available from: https://arxiv.org/pdf/1709.02925.pdf [Accessed 30/01/2018]*.

## References

1. Brownlee J. (2014) A Data-Driven Approach to Choosing Machine Learning Algorithms. [Online] September 29th 2014. Available from: Machinelearningmastery.com https://machinelearningmastery.com/a-data-driven-approach-to-machine-learning/. [Accessed: 30 January 2018].
2. Ren S., Liao B., Zhu W., Li K. (2018) Knowledge-maximized ensemble algorithm for different types of concept drift. Information Sciences. 430–431 (3). P. 261–281.
3. Hsu C-H., Shi X., Salapura V. (2014) 11th IFIP International Conference on Network and Parallel Computing (NPC), 18-20 September 2014. Ilan, Taiwan. Springer. LNCS-8707. P. 333–346, 2014. Network and Parallel Computing.
4. Diez-Pastor J.F., Rodríguez J.J., García-Osorio C. & Kuncheva L.I. (2015) Diversity techniques improve the performance of the best imbalance learning ensembles. Information Sciences. 325. P. 98–117.
5. Löfström T. (2015) On Effectively Creating Ensembles of Classifiers: Studies on Creation Strategies, Diversity and Predicting with Confidence. A Thesis Submitted in partial fulfilment of the Requirements of Stockholm University for the Degree of Doctor of Philosophy. Stoke-on-Trent: Stockholm University.
6. Hamed R.B., Fazli C. (2017). Less Is More: A Comprehensive Framework for the Number of Components of Ensemble Classifiers. IEEE Transactions on Neural Networks and Learning Systems. [Online] 14(8), September 2017 USA: IEEE. P. 1–7. *Available from: https://arxiv.org/pdf/1709.02925.pdf [Accessed 30/01/2018]*.

***Про авторів:***

[1]*Крак Юрій Васильович*,
доктор фізико-математичних наук, професор,
завідувач кафедри теоретичної кібернетики Київського національного університету імені Тараса Шевченка,
старший науковий співробітник Інституту кібернетики імені В.М. Глушкова НАН України.
Кількість друкованих праць – понад 500, в тому числі:
кількість наукових публікацій в українських фахових виданнях – 170,
кількість наукових публікацій в зарубіжних виданнях – 60.
Н-індекс – 2.
http://orcid.org/0000-0002-8043-0785,

[2]*Бармак Олександр Володимирович,*
доктор технічних наук, професор,
професор кафедри Комп'ютерних наук та інформаційних технологій
Хмельницького національного університету.
Кількість друкованих праць – понад 200, в тому числі:
кількість наукових публікацій в українських фахових виданнях – 70,
кількість наукових публікацій в зарубіжних виданнях – 15.
Н-індекс – 1.
http://orcid.org/0000-0003-0739-9678,

[2]*Манзюк Едуард Андрійович*,
кандидат технічних наук, доцент кафедри Комп'ютерних наук та інформаційних технологій
Хмельницького національного університету.
Кількість друкованих праць – понад 30, в тому числі:
кількість наукових публікацій в українських фахових виданнях – 20.
http://orcid.org/0000-0002-7310-2126.

***Місце роботи авторів:***

[1]Київський національний університет імені Тараса Шевченка,
01601, Київ, вул. Володимирська, 60.
E-mail: krak@unicyb.kiev.ua,
yuri.krak@gmail.com,

[2]Хмельницький національний університет МОН України,
29016, Хмельницький, вул. Інститутська, 11.
E-mail: alexander.barmak@gmail.com,
eduard.em.km@gmail.com