Формування Масиву Вхідних Даних для Мамографічних Досліджень

С. Палаш
Кафедра інтелектуальних систем прийняття рішень
Черкаський національний університет імені Богдана Хмельницького
Черкаси, Україна
s_palash@ukr.net

О. Харченко
Кафедра інтелектуальних систем прийняття рішень
Черкаський національний університет ім. Богдана Хмельницького
Черкаси, Україна
alexandrharchenko106@gmail.com

Formation of an Array of Input in the Mammographic Research

S. Palash
Department of intelligent systems of acceptance of decision
Cherkasy National UniversityBohdan Khmelnytsky
Cherkasy, Ukraine
s_palash@ukr.net

A.Kharchenko
Department of intelligent systems of acceptance of decision
Cherkasy National University Bohdan Khmelnytsky
Cherkasy, Ukraine
alexandrharchenko106@gmail.com

***Анотація*—представлено порівняння методів формування масиву вхідних даних для дослідження мамографічних знімків. Експерементально підтверджено кращі результати розпізнавання змішаним методом формування масиву вхідних даних.**

***Abstract*—presented a comparison of methods for forming an arrays of input data for the mammography image analysis. Experimentally proved that the better method for mammography recognition and analysis is a mixed method.**

Ключові слова—масив вхідних даних, мамограмирозпізнавнання.

Keywords—arrays of input data, mammography image, recognition.

I. Вступ

Аналіз мамографічних зображень актуальна задача, яка потребує постійного вдосконалення. Для дослідження мамограм, формуються масиви вхідних даних. Встановивши вид матриці вихідних даних, приступають до формування інформаційного масиву. На цьому етапі визначають перелік змінних і об'єктів спостереження. Відбір ознак (змінних)і об'єктів спостереження є дуже важливим етапом роботи і виконується в суворій відповідності з метою дослідження. Якість інформації тут значною мірою залежить від знань про об'єкт і ознак, які мають найбільш істотну інформацію про цей об'єкт. Важливими критеріями відбору при формуванні масиву вхідних даних (МВД) є інформативність показників і точність відображення їх чисельних характеристик.

II. Постановка задач досліджень

Метою даної роботи є дослідження процесу формування МВД в умовах різної інформативності показників. Передбачається досягнути мету за рахунок використання різних алгоритмів збору даних. Мета буде досягнута тоді, коли інформативність МВД буде достатньою для синтезу корисних моделей. Моделі повинні розв'язати задачу аналізу знімків.

III. Формування масиву вхідних даних

Масив вхідних даних – це матриця чисельних характеристик елементів зображення, отриманих в процесі його декомпозиції.

Процес формування МВД має принципово важливе значення для успішного виконання завдань машинного навчання. Часто завдання машинного навчання виконуються саме завдяки правильного формування масиву даних для навчання моделей – навчальної множини. Помилки у формуванні навчаючої множини зазвичай виявляються критичними і здатні звести нанівець ефективність самих алгоритмів навчання. Серед фахівців по машинному навчання загально визнаним вважається, що наявність хороших навчальних даних набагато важливіше якості алгоритму навчання.

У зв'язку з активним розвитком систем багаторівневого перетворення інформації в останнє десятиліття питання формування множини навчальних даних приймають особливо важливе значення, оскільки в багатьох задачах система багаторівневого перетворення даних демонструює якість, істотне покращення ніж інші алгоритми машинного навчання, однак, щоб отримати подібний виграш як, необхідно використовувати навчальну множину дуже великого розміру (до декількох мільйонів зображень, при цьому навчання вимагає великого обсягу обчислювальних ресурсів і може займати кілька тижнів на багатопроцесорному кластері).

Використання методу групового урахування аргументів дає кращий результат в умовах обмеженої інформативності.

Представлення навчаючої вибірки, можна описати так:

Нехай маємо множину об'єктів G, множина відповідей P, і існує цільова функція p^*: G → P, значення якої $P_i = p^* * (g_i)$ відомі тільки для підмножини об'єктів $\{1, .., g_k\} \subset G$. Сукупність пар $G^K = (g_i, p_i)_{i=1}^k$ називається вибіркою. Задача навчання полягає в тому, щоб по навчаючій вибірці G^k відновити залежність p^*, тобто побудувати вирішуючу функцію a: G → P, яка б наближувала б функцію $p^*(g)$, причому не тільки на навчаючій вибірці, але ні на всій вибірці G [1]. Метод навчання – це відображення μ: $(G \times P)^k \to A$, яке у довільній конечній навчаючій вибірці $G^k = (g_i, p_i)_{i=1}^k$ ставить у відповідність вирішуючу функцію a: G → P. Метод навчання μ будує вирішуючу функцію a по навчаючій вибірці G^k [1].

Правильне формування навчальної вибірки часто має вирішальне значення в задачах машинного навчання, що визнається більшістю фахівців в даній області. Найчастіше рішення задач машинного навчання зводиться до грамотного формування навчальної вибірки. Незважаючи на це, в сучасній літературі по машинному навчання питань формування навчальної вибірки майже не приділяється увага, теоретична база практично відсутня.[1].

Способи генерації навчальної множини описані нижче.

Програмна генерація - генерація навчальних даних по деякому алгоритму. Алгоритм генерації і його параметри, в даному випадку, визначають отримані на виході розподілення у просторі об'єктів. Доцільно варіювати якомога більше параметрів в процесі генерації. Однак метод може виявитись неефективним та згенерувати велику кількість не потрібних даних.

Якщо реально можлива лише мала частина об'єктів з простору параметрів генерації або найбільш типові об'єкти мають складну конфігурацію.

Спосіб покликаний подолати недоліки попереднього методу – семплірування. Алгоритм намагається згенерувати вибірку з даного розподілу, здається деякий апріорний розподіл в просторі об'єктів. Семплірування по Monte Carlo, Гиббсу, схема Метрополіс – Гастингса, застосовуються методи вибірки з відхиленням. Застосування даних методів використовується для того, щоб досліджувати найбільш осмислені частини простору об'єктів (зазвичай це лише мала частина потенційно можливих об'єктів, коли простий перебір об'єктів з перевіркою їх коректності може зайняти невиправдано багато часу).

Закономірна модифікація базового об'єкта. В даному випадку маємо набір базових об'єктів, навчальна множина формується шляхом багаторазової модифікації їх параметрів (як правило, зовнішніх змінних). Прикладом може бути вибірка кадрів з відеопослідовності, вибірка поз людини з mocap-а. Метод доцільно застосовувати в тих випадках, коли немає можливості аналітично задати розподіл можливих об'єктів в просторі об'єктів, або не підходить використання синтетичних даних. При використанні даного методу генерації навчальної множини зберігається велика кількість фонових закономірностей. Крім того, виникають проблеми, якщо отриману таким чином множину розбити в деякому відношенні на навчальну і тестову і використати отриману тестову множину для контролю за перенавчанням. Оскільки об'єкти результуючого мають сильну взаємозалежність, низький рівень помилки алгоритму на тестовій вибірці не гарантує відсутності «заучування» навчальних даних. При використанні метричних алгоритмів класифікації найефективніше в даному випадку - завжди відносити об'єкт до того самого класу, до якого належить найближчий навчальний об'єкт. Приклад помилки подібного роду при використанні алгоритму k-NN наведено в роботі [4] (при налаштуванні гіперпараметрів алгоритму методом ковзаючого контролю за багатьма даних, отриманого вищеописаним методом, завжди вибирався параметр k = 1). Тому при використанні даного методу формування множини даних не варто застосовувати метод ковзаючого контролю в чистому вигляді.

Ще один спосіб – вибірка з бази об'єктів. Формування множини зображень з довільної множини зображень обличь людини погано описуються наведеними вище моделями. Є фіксований набір об'єктів із заздалегідь визначеними параметрами і ми не можемо отримати об'єкт з будь-якими заданими внутрішніми параметрами. Якщо це зображення місцевості, розташування цих об'єктів підпорядковується строгим обмеженням. Об'єкти можна розбити на групи, причому об'єкти всередині групи будуть дуже схожі, а об'єкти з різних груп - відрізнятися (можна порівняти зображення осіб злочинців і зображення облич політиків). Об'єкти, географічно розташовані близько один до одного, зазвичай більш схожі, ніж об'єкти, що знаходяться далеко один від одного (наприклад, будинки в одному і в різних містах). Якщо це колекція фотографій, слід звертати увагу на переваги фотографа, тому, що люди набагато частіше фотографують красиві речі (наприклад, пам'ятки), ніж одноманітний ліс або пустелю, різні люди роблять знімки різних типів.

При генерації навчальної множини даними способом – важко гарантувати наявність усіх принципово важливих типів об'єктів в множини даних. В даному випадку може допомогти перехід до автоматичної генерації. Також слід

робити вибір об'єктів з якомога більш широкого набору груп об'єктів або місць зйомки (детектор особи, навчений на базі осіб злочинців, можливо, буде не дуже добре працювати в загальному випадку).

IV. Способи додавання даних в навчальну множину

Додавання даних є одним з найпростіших і ефективних способів поліпшити якість навчальної множини. При цьому, незавжди ефективно просте додавання даних довільного виду часто потрібно додати дані певного різновиду для підняття якості розпізнавання. Розглянемо деякі способи додавання даних.

Програмна генерація. У разі використання синтетичних навчальних даних зручніше всього згенерувати відсутні навчальні приклади. Однак не у всіх завданнях допустимо використання програмно згенеровані дані. У таких випадках доводиться застосовувати складніші методи додавання даних.

Data augmentation. Модифікація наявних зображень з метою розширити навчальну вибірку. Активно застосовується в системах багаторівневого перетворення даних [5], а також в умовах дефіциту розмічених даних. Застосовуються стиснення/розтягування, горизонтальне відображення, поворот, випадковий зсув в колірному просторі, випадкове або закономірна зміна деяких пікселів. Вважається, що додавання повністю випадкового шуму неефективно, слід додавати шум, обумовлений даними (тільки потенційно можливі в реальних даних спотворення). Істотний недолік даного методу - зберігається більшість фонових закономірностей.

Hard samples mining [6]. Класична проблема в задачах пошуку об'єктів на зображенні - потреба в підтримці достатнього числа hard negative samples (навчальних прикладів, що схожі на об'єкт інтересу, але такими не є) в навчальній множині. Складність виникає через те, що в природних умовах такі об'єкти зустрічаються рідко, тому застосовуються спеціальні методи для їх пошуку і додавання в навчальну множину (hard samples mining). Ключове припущення в даних методах - цікавлять нас об'єкти позитивних але схожі між собою. Зазвичай застосовуються data augmentation, адаптивний пошук, пошук по шаблонах, методи на основі машинного навчання. Цікавим є застосування методів тематичного моделювання для пошуку складних негативних прикладів.

Імітація додавання даних. При навчанні систем багаторівневого перетворення інформації можливе використання методу dropout [6]: випадкове обнулення активацій деяких моделей в мережі при подачі їй на вхід чергового тренувального зображення (зазвичай в кожному шарі випадково вибирається 20-50% моделей). Без застосування даної техніки система багаторівневого перетворення інформації може «заучувати» велику кількість фонових закономірностей через те, що складність моделі перевищує обсяг доступних даних. По суті, dropout - це імітація додавання даних в навчальну вибірку. Усередині алгоритму навчання ми імітуємо мінливість даних - на вхід більш глибоких рівнів мережі надходить випадковим чином змінена версія реального зображення (мається на увазі зображення, що не викликало б активацію обнулення моделей), хоча таких даних насправді немає в навчальній множині. Недолік даного методу - може бути імітований додавання таких даних, яких в принципі не може бути в реальності, через чого може страждати точність розпізнавання. Цікавим є створення модифікацій даного методу, що враховують природу даних.

Краудсорсинг. Оскільки для систем багаторівневого перетворення інформації потрібні величезні об'єми вручну розмічених навчальних даних, для формування навчальної множини активно використовуються сервіси краудсорсингу - користувачі сервісу за невелику плату створюють розмітку «сирих» даних (наприклад, вказують, які об'єкти є на даному зображенні і де вони розташовані). Найпопулярніший з таких сервісів - Amazon Mechanical Turk [7].

Проблема даного методу - велика кількість помилок в розмітці, так як користувачі не завжди роблять свою роботу сумлінно, а іноді просто помиляються, тому потрібні спеціальні методи контролю помилок в розмітці. Зазвичай одну і ту ж картинку дають розмітити декільком користувачам, а потім вибирають той варіант розмітки, який вибрало найбільше число користувачів. Також вводяться спеціальні метрики «сумлінності» користувача.

V. Результати роботи алгоритмів формування МВД

Для проведення аналізу формування МВД використовувалися такі методи:

Ручне формування МВД.

Програмна генерація МВД.

Data augmentation.

Hard samples mining.

Формування МВД шляхом ковзаючого вікна.

Змішані методи.

В таблиці 1 представлені порівняння результатів роботи даних методів та показники стійкості та адекватності моделей. Для тестування було відібрано зображення розмічене експертом. В результаті експерт виявив 10 ділянок з підозрою на ущільнення тканин, 2 ділянки з мікрокальцинатами та 2 ділянки з підозрілою поведінкою. Показники стійкості та адекватності визначалися в межах від 0 до 1, де 0 – абсолютно стійка модель, тобто та, що розпізнає образ в умовах зашумленості зображення. За даною таблицею легко побачити, що формування масивів вхідних даних змішаним методом дає найкращий результат.

ТАБЛИЦЯ І. ПОРІВНЯННЯ МЕТОДІВ

Назва методу	*Кількість розпізнаних зображень*	*Стійкість*	*Адекватність*
Ручне формування МВД	8	0,7	0,75
Програмна генерація МВД	8	0,72	0,75
Data augmentation	9	0,8	0,7
Hard samples mining	10	0,82	0,72
Формування МВД шляхом ковзаючого вікна	10	0,85	0,76
Змішані методи	13	0,9	0,9

Нижче наведено результати роботи програми у вигляді графічних зображень.

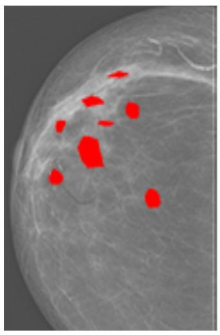

Рис. 1. Результати роботи методу Hard samples mining

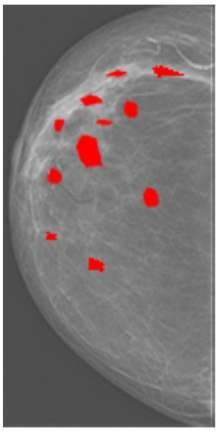

Рис. 2. Результати роботи методомко взаючого вікна

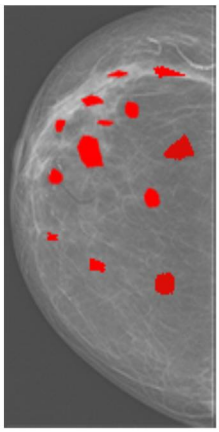

Рис. 3. Результати роботи змішаного методу

Змішаний метод формування масиву вхідних даних дав кращий результат. Оскільки було розпізнано 13 зображень із 14.

VI. ВИСНОВОК

Дана робота демонструє способи формування масиву вхідних даних різними методамби. Результати досліджень дозволяють зробити висновок, що формування масиву вхідних даних змішаними методами показують кращий результат. Даним методом було вірно розпізнано 93% ділянок.

ЛІТЕРАТУРА REFERENCE

[1] И.Л. Кафтанников, А.В. Парасич Проблемы формирования обучающей выборки в задачах машинного обучения Южно-Уральский государственный университет, г. Челябинск, Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2016. Т. 16, № 3. С. 15–24

[2] Mangalova E., Petrun'kina I. [Prediction Capacity of Wind Power Plants Based on NonParametric Algorithm, K Nearest to Neighbors]. Doklady vserossiyskoy nauchnoy konferentsii AIST'2013 [Reports of the All-Russian Scientific Conference AIST'2013]. Ekaterinburg, 2013, pp. 1–8. (in Russ.) J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] Canavet O., Fleuret F. Efficient Sample Mining for Object Detection. Proceedings of the Asian Conference on Machine Learning (ACML), 2014, pp. 48–63.

[4] . Srivastava N. Hinton G.E., Krizhevsky A., Sutskever I., Salakhutdinov R.R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. The Journal of Machine Learning Research, 2014, vol. 15, no. 1, pp. 1929–1958

[5] Голуб С.В. Багаторівневе моделювання в технологіях моніторингу оточуючого середовища / С.В. Голуб. – Черкаси: Вид. від. ЧНУ імені Богдана Хмельницького, 2007. – С.220.

[6] Amazon Mechanical Turk. Avaliable at: https://www.mturk.com/mturk/welcome (accessed September 2015).

[7] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.