UNITECH '10 GABROVO

INTERNATIONAL SCIENTIFIC CONFERENCE
19 – 20 November 2010, GABROVO

ПАРАЛЕЛЕН АЛГОРИТЪМ ЗА РАЗПОЗНАВАНЕ НА ПЛАСТМАСИ С ИЗПОЛЗВАНЕ НА УЕЙВЛИТНИ ТРАНСФОРМАЦИИ

PARALLEL ALGORITHM FOR PLASTICS IDENTIFICATION USING WAVELET TRANSFORMATIONS

Пламенка Боровска
ФКСУ, Технически университет - София

Десислава Иванова
ФКСУ, Технически университет - София

Abstract

The paper presents realization of a fast parallel algorithm for training and classification of plastic specimens by using k-NN classifier, and signs result from factors wavelet decomposition of the input signal. The chosen method of parallelization is based on a hybrid model, a combination of multithreaded processing and message passing interface (MPI). Multi-threaded processing is done using the standard directives OpenMP, and for communication between processes is used message passing interface MPI (in the implementation MPICH2). Measurements are conducted for estimation of speed up and efficiency as a function of the core numbers, the number of parallel processed samples, the sample size and the different types of parallelization as a function of the number of samples.

Keywords: Mallat wavelet transformation, synthesis classifier, spectral analysis, parallel code (MPI; OpenMP), parallel program implementation, evaluation of performance parameters

ВЪВЕДЕНИЕ

Трудно можем да си представим съвременния свят без пластмасата – този толкова пластичен, устойчив и траен материал. Около 50 различни вида пластмаса се използват, за да се изработят предмети, всеки от които има различен химичен състав и редица свойства, които го правят приложим за специфични дейности. Разграждането им в природата обаче е един много дълъг процес. Тъй като пластмасите биват различни по състав, те задължително се сортират преди да се рециклират в нови продукти. Това се явява проблем! Голямото им разнообразие прави трудно разпознаването им. Нужни са мощни методи и алгоритми и добри технически средства за разработването на система за безконтактна идентификация на материалите, които да дават бърза и точна оценка на принадлежността им към определен клас.

Чрез съвременните технологии за повишаване на производителността на компютърните системи се постигат все по голям брой изчислителни елементи. Паралелните компютърни архитектури стават все по-достъпни и масови, а изчислителната мощ нараства все повече. Такъв е примерът с разработването на автоматична система за безконтактна класификация на пластмаси, даваща обективна, точна и бърза информация за тяхната принадлежност.

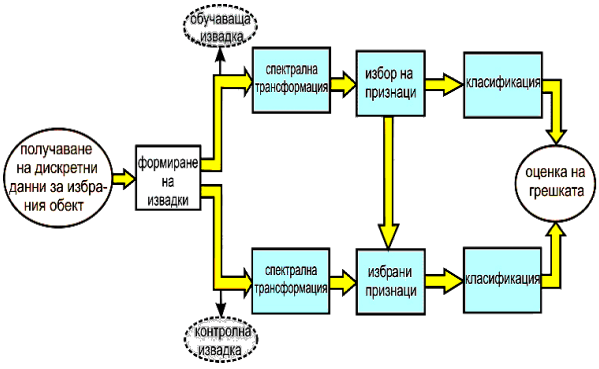
В статията са представени резултати от изследване на автоматични класификатори – ядрата на сортираща машина за пластмаса, определящи в най-голяма степен тяхната физическа реализуемост и ефективност. Насоката на изследванията е в ускоряване на работата на класификатора посредством паралелно изпълнение на елементи от алгоритъма за обучение. Изследваните класификатори са базирани на спектрален анализ на входните сигнали посредством дискретно уейвлитно преобразуване (DWT).

СПЕКТРАЛЕН АНАЛИЗ С ИЗПОЛЗВАНЕ НА УЕЙВЛИТНИ ПРЕОБРАЗУВАНИЯ

При анализ и обработка на нестационарни във времето или неоднородни в прос-

трансството сигнали, освен общата честотна характеристика на сигнала, са необходими и сведения за определени локални координати, в които произтичат изменения на честотните съставящи на сигнала. Получава се така наречената "спектрограма", при която не е възможно да се определи точно времечестотно представяне на сигнала. Времечестотното представяне на сигнала става възможно, ако като базисни функции се използват компактни вълнови функции – т.нар. уейвлити.

- *Синтез на система за автоматична класификация на пластмаси чрез използване на метода за спектрален анализ на сигнали* - задачата за синтез на система за автоматична класификация се свежда до задача за анализ и разпознаване на отразени ултразвукови сигнали. Процесът се състои в преобразуване на информацията за стойностите на признаците, описващи обектите за класифициране, в информация за принадлежността им към предварително определен клас.
- *Формиране на признаково пространство с използването на уейвлит преобразувания* - За да бъде приложимо дадено преобразуване за разпознаване на пластмаси, то трябва да е базирано на ортогонална система базисни функции и да съществуват бързи алгоритми за преобразуване. Подходящо е дискрет-ното уейвлитно преобразуване (DWT) с ортогонални уейвлити с разработеното бързо уейвлитно преобразуване (алго-ритъм на Mallat). За разлика от спект-ралното представяне на Фурие, уейвлит спектърът е двумерен и образува дву-мерна повърхност на променливите a и b.
- *Избор на уейвлитна базисни функция* - Изборът на конкретен уейвлит е обусловен от практическото приложение и типа на сигнала - гладкост, необходимо ниво на разлагане, нарастващо време за трансформация с увеличаване порядъка на уейвлита и др. [8,9,10]. Могат да се разграничат няколко по-важни групи уейвлити: Предуейвлити - гаусови уейвлити, Морле, мексиканска шапка (mexihat); Уейвлити на Майер - регулярни (meyr) и дискретни (dmey); Ортогонални уейвлити с компактен носител – уейвлити на Добеши (dbN), симлети (symN), коифлети (coifN); Биортогонални уейвлити с компактен носител.
- *Синтез на класификатори* - Общата структура на задачата “синтезиране на класификатор” за безконтактно разпознаване на пластмаси с приложение на спектрални методи за формиране на признаци е представена на фиг. 1.



Фиг. 1. *Последователност на процеса синтез на класификатор.*

От изходните данни са формирани обучаваща и контролна извадка. С помощта на бързи уейвлитни трансформации са формирани признаци и е извършен подбор по метода на последователното изключване. С подбраните признаци е оценена точността посредством класификация на контролната извадка. Пригодността на формираните признаци е изследвана с класификатори, използващи методите на “к-най-близки съседа” за разпознаване на образи [1], като при реализация се определят тегловните коефициенти:

$$\lambda_j^{(i)} = \frac{1}{\sigma_j^{(i)} \sum_{j=1}^{m} \frac{1}{\sigma_j^{(i)}}} \quad ; \quad \sum_{j=1}^{m} \lambda_j^{(i)} = 1 \qquad (1)$$

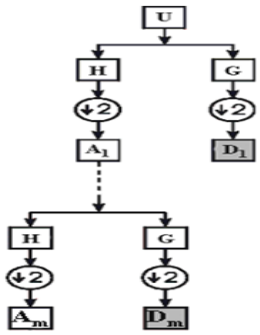
$$\lambda_j^{(i)} = \frac{1}{\sigma_j^{(i)} \sqrt[m]{\prod_{j=1}^{m} \frac{1}{\sigma_j^{(i)}}}} \quad ; \quad \prod_{j=1}^{m} \lambda_j^{(i)} = 1 \qquad (2)$$

където: σ_j е средно-квадратичното отклонение на всеки признак; i – клас на обектите; j – номер на съответния признак.

ПАРАЛЕЛЕН АЛГОРИТЪМ ЗА РЕАЛИЗИРАНЕ НА УЕЙВЛИТ ТРАНСФОРМАЦИЯ

Неделима част от създаването на алгоритъм за реализиране на уейвлит трансформация е изборът на начин за паралелизацията му. Целта е да бъде постигната висока производителност и минимална грешка при класификацията.

1) Пирамидален алгоритъм на Маллат (Mallat) - За ортогоналните уейвлити съществува разработено бързо уейвлитно преобразуване, известно още като пирамидален алгоритъм на Маллат (Mallat) [1, 2]. При този алгоритъм трансформацията се извършва под формата на серия от линейни филтърни операции, което силно опростява практическото реализиране на уейвлит преобразуването.



Фиг. 2. *Схема на бързо уейвлитно преобразуване на m нива*

сигналът U се подава на нискочестотен филтър H и високочестотен G с предавателни функции, съответно:

$$H(\omega) = \sum_{n \in Z} h_n e^{-in\omega} \qquad (3)$$

$$G(\omega) = \sum_{n \in Z} g_n e^{-in\omega} \qquad (4)$$

съответстващи на уейвлитите $\psi(t)$ и $\varphi(t)$. Тук h_n и g_n са коефициенти на филтрите и се изчисляват в зависимост от прилагания уейвлит, а n е цяло число. След операцията двоична децимация ($\downarrow 2$), с която броя на честотните съставящи намалява на половина, от филтър H се получават коефициентите на апроксимация $\mathbf{A_1}$ на ниво m=1, а от филтър G – детайлизиращите коефициенти $\mathbf{D_1}$.

2) Избор на подходящо ниво на разлагане m (декомпозиция) - Изборът на подходящо ниво на разлагане (m) е от съществено значение за правилния синтез на признаци за класификация. Апроксимиращите и детайлизиращите уейвлитни коефициенти, характерни за ортогоналните уейвлити, трябва достатъчно точно да представят изходните реализации, като отразяват особеностите, характерни за класифицираните обекти. Най-добрият базис представя изходната реализация с най-малък брой значими коефициенти, отчитащи характерните особености на всеки от класовете.

Известно е, че средното количество информация, която се съдържа в една реализация, се изразява с ентропията на сигнала:

$$H(U) = -\sum_k H(U_k) = -\sum_k U^2(k) \log_2(U^2(k)) \quad (5)$$

Ентропията е основен параметър, характеризиращ сигнала и предоставя възможност за избор на подходящо ниво на разлагане. Използва се базисът, който предоставя най-малка ентропия.

3) Паралелни алгоритми и начини за разработка - Съвременните среди за паралелно програмиране са от типа MIMD [3]. Те могат да се класифицират според вида на използвания паралелизъм.

- SPMD (Single Program Multiple Data) – оползотворяват се паралелизми по данни. Кодовете на паралелните процеси са еднакви, но всеки процес обработва различни данни.
- MPMD (Multiple Program Multiple Data) – оползотворява се функционален паралелизъм. Процесите са хетерогенни т.е. всеки процес изпълнява различен код над различни данни.

В рамките на една и съща програма могат да се комбинират и двата стила на паралелно програмиране.

Систематизирането на алгоритмите и приложенията трябва да се основава на фактори, които напълно характеризират паралелния алгоритъм, като свойства на процесите (структура, топология и изпълнение), свойства на взаимодействие и качества на данните (разделяне и разположение). Най-популярните парадигми при паралелното програмиране могат да бъдат определени както следва:

– Парадигма “Една програма множествени данни” (SPMD);

– Парадигма “Разделяй и владей”;

– Парадигма „Главен подчинен”;
– Парадигма „Конвейеризация на паралелните изчисления”;
– Парадигма „Работен пул”;
– „Фазово паралелна”;
– Парадигма на „Синхронните и асинхронните итерации”.

В програмната реализация се използват парадигмите „Главен подчинен” и „Работен пул”.

Парадигмата „Главен подчинен” е известна още като “процесорни ферми”. При нея главният процес изпълнява основната последователна част на програмата и разпростира множеството подчинени процеси, обработващи паралелния работен товар. Когато някой от подчинените процеси завърши изпълнението на текущата задача, той информира главния процес, който от своя страна му задава за изпълнение нова задача. Тук координацията се осъществява от главния процес, а комуникацията най-често се осъществява между главния процес и подчинените процеси.

Парадигмата „Работен пул”, при нея се използва моделът с общи променливи. Пулът съдържащ работния товар, се реализира като глобална структура от данни. Създава се множество паралелни процеси. Първоначално в работния пул може да има само един изчислителен товар. Всеки свободен процес може да взима произволен товар от пула и да го изпълни, като в резултат на това изпълнение може да генерира допълнителни единици изчислителен товар и да ги добави в работния пул. Изпълнението на паралелната програма завършва когато се изчерпи изчислителният товар в работния пул. Тази алгоритмична парадигма осигурява изключително добро балансиране на товара, тъй като работният товар динамично се разпределя между свободните процеси.

В зависимост от типа на използвания паралелизъм, паралелните програмни модели се разделят на два типа:

- С неявен паралелизъл – за оползотворяването му се използва автоматична паралелизация на последователните програми. За целта се използват компилатори, които правят анализ на зависимостите по данни и управление в изходния код на последователната програма и след това прилагат различни техники за трансформиране на кода с цел да може да се паралелизира. Тези компилатори се наричат паралелизиращи компилатори.
- С явен паралелизъм - предпочитани поради ниската ефективност на автоматичната парализация.

Най-популярните програмни модели с явен паралелизъм са:

– моделът с паралелизми по данни – използва се при режими SIMD или SPMD. При тях едновременно се изпълнява един и същ код над различни потоци данни в различни изчислителни възли. При този стил на паралелно програмиране управляващият поток е един, а паралелизмите се оползотворяват на ниво данни.

– моделът с обмен на съобщения - (MPI - Message Passing Interface);

– паралелни програмни модели с обща памет – (OpenMP).

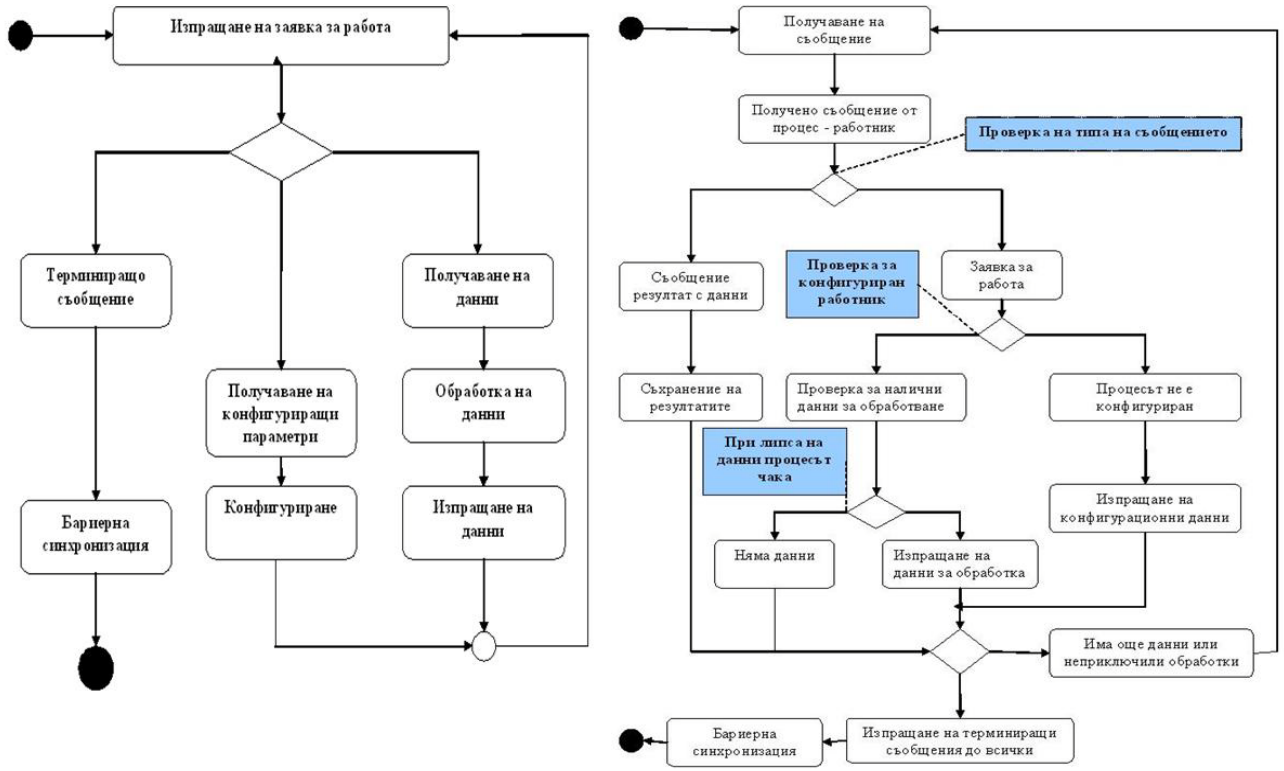
РАЗРАБОТВАНЕ НА ПАРАЛЕЛНИ ПРОГРАМНИ ИМПЛЕМЕНТАЦИИ

Програмата реализира модел, в който процесите се делят логически на два типа. Един от всички процеси, а именно този с ранг 0, изпълнява ролята на главен процес и управлява действието на останалите процеси, които в случая изпълняват ролята на процеси-работници. При първото поискване на работа от всеки един процес-работник, вместо работна единица той получава съобщение с конфигурационна информация (с маркер TAG_CONFIGURE), включваща типа на използвания уейвлит, нивото на декомпозиция, както и списък на коефициентите, изграждащи множеството от признаци.

Процесите-работници имат проста логика. В момент, в който са свободни, поискват работна единица от главния процес, като изпращат празно съобщение със съответен маркер (TAG_READY). В резултат могат да получат или съобщение с конфигурационна информация (с маркер TAG_CONFIGURE), или съобщение с данни за обработка (с маркер TAG _WORK), или терминиращо съобщение (с маркер TAG_EXIT).

Въпреки, че наличието на главен процес

и процеси-работници причислява алгоритъма към реализациите на метода «главен-подчинен», то на практика се постига реализация на «Работен пул».



а) *б)*

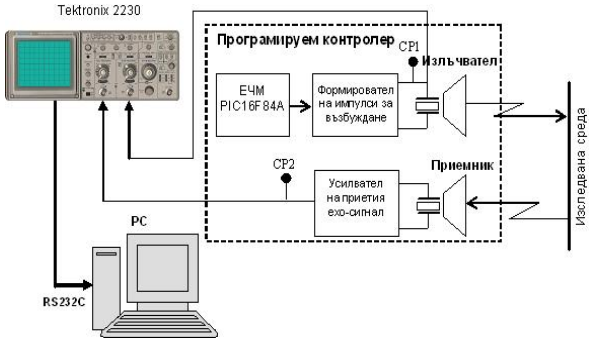
Фиг. 3. *Диаграма на действие на процес-работник (а) и диаграма на действие на главен процес (б)*

ЕКСПЕРИМЕНТАЛНА ПОСТАНОВКА

Проведените експерименти за получаване на информация са реализирани в off-line режим с помощта на ултразвукови датчици тип UST40T/UST40R на фирмата Nippon Ceramic company. На базата на анализ на върнатия след отражение ултразвуков сигнал могат да се разпознават различни физични среди, като предварително са определени отличителните им характеристики. Експерименталната схема на опитната постановка е показана на фиг. 4, която се състои от апаратни и програмни модули. Апаратните модули включват:

- PC конфигурация с Microsoft Windows XP.
- Програмируем контролер на базата на едночипов микроконтролер (ЕЧМ) PIC16F84A.
- Запомнящ осцилоскоп Tektronix 2230.

Програмните модули се състоят от програма за управление на програмируемия контролер, който генерира импулсно модулиран сигнал за възбуждане на излъчвателя UST40T. Отразеният сигнал се записва от осцилоскопа, след което се прехвърля в компютъра по сериен интерфейс RS-232C.



Фиг. 4. *Схема на експерименталната постановка*

АНАЛИЗ НА ЕКСПЕРИМЕНТАЛНИТЕ РЕЗУЛТАТИ И ОЦЕНКА НА ПАРАМЕТРИТЕ ЗА ПРОИЗВОДИТЕЛНОСТ

Реализираната паралелна програмна имплементация се изпълнява на клъстера на

ТУ-София, който се състои от общо десет възела, които могат да бъдат класифицирани по следния начин:

- Конфигурация на възел тип 1: два четириядрени процесора Intel® Xeon® E5405 и 4 GB RAM памет.
- Конфигурация на възел тип 2: двуядрен процесор AMD Opteron 1210 и 2 GB RAM памет.

Клъстерът се състои от 2 възела от тип 1 и 8 възела от тип 2.

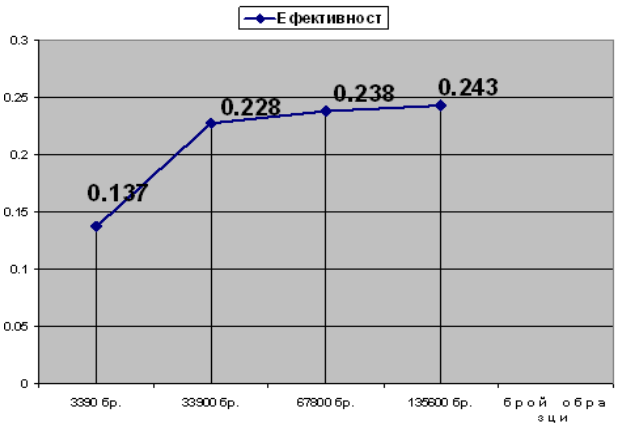
Изследвана е ефективността на програмата в случая, когато всеки процес-работник използва толкова нишки, колкото процесорни ядра има машината, на която се изпълнява – общо 9 процеса и 30 нишки и 1 нишка за главния процес. Като параметър на изследването е приет броят на обработваните образци, ефективно управляван посредством манипулиране на броя циклични обработки на входната извадка от 339 образеца. Ускорението и ефективността са изчислени спрямо времето, за което последователната програма се изпълнява на машина с процесор Opteron.

Брой образци	Време за изп. последователно	Време за изп. 31 нишки	Ускорение	Ефективност
3390	5.246 s	1.233 s	4.255	0.137
33900	46.643 s	6.591 s	7.077	0.228
67800	92.661 s	12.545 s	7.386	0.238
135600	184.716 s	24.458 s	7.552	0.243

С увеличаването на броя на образците ускорението нараства, като достига стойност от порядъка на 7.55 при обработването на 135600 образци (фиг.5).



Фиг. 5. *Ускорение при различен брой образци*



Фиг. 6. *Ефективност при различен брой образци*

С увеличаването на броя на образците ефективността нараства (фиг.6), като се наблюдава тенденция към достигане на гранична стойност от порядъка на 0.25. Ниската ефективност се дължи на относително големия дял на комуникацията (MPI) и синхронизацията (при OpenMP) спрямо дяла на фактически извършваната работа.

Горна граница на ускорението поставя способността на свързващата мрежа да предаде пакетите с данни за преобразуване до процесите-работници. Един пакет от 2560 дискрета като числа с плаваща запетая и двойна точност има размер 20480 байта (2560 дискрета * 8 байта). Дори ако се пренебрегнат всички протоколни свръхтовари, то 1 Gbps Ethernet не може да предаде повече от $10^9/(20480*8) = 6103$ образеца/ секунда.

ЗАКЛЮЧЕНИЕ

В статията бяха представени резултатите от проектиран и имплементиран бърз паралелен алгоритъм за обучение и класификация на образци от пластмаса посредством използване на k-NN класификатор. Избран е метод на паралелизация, базиран на хибриден модел, комбинация от многонишкова обработка и предаване на съобщения. Направена е паралелна програмна имплементация на езика C. Реализиран е многонишков вариант на паралелния алгоритъм на Маллат за дискретно уейвлит преобразуване. Многонишковата обработка е направена с използване на директиви от стандарта OpenMP, а за предаване на

съобщения между отделните процеси е използван MPI (в реализация MPICH2).

ПРИЗНАТЕЛНОСТ

Докладът е финансиран по договор № ДРНФ 02/9 – 15.12.2009 г.

ЛИТЕРАТУРА

[1] S. Mallat, ‘‘Multiresolution approximations and wavelet orthogonormal bases of L2(R),’’ Trans. Amer, Math. Soc., 315:69--87, 1989

[2] Дьяконов Вл., Вейвлеты. От теории к практике. Солон-Р, Москва, 2002

[3] Боровска П., Лазарова М., Паралелна информационна обработка, CIELA, София 2007