ПАРАЛЕЛЕН EM-АЛГОРИТЪМ ЗА СМЕСЕНО МАШИННО САМООБУЧЕНИЕ (MT-SSEM)

Гергана Лазарова[1], Иван Койчев[1,2]

[1]Факултет по математика и информатика,
Софийски университет „Св. Климент. Охридски”
[2]Институт по математика и информатика - БАН
gerganal@fmi.uni-sofia.bg, koychev@fmi.uni-sofia.bg

***Резюме**: С развитието на информационните технологиите събирането на огромно количество данни става все по-лесно. Натрупани са големи масиви от данни и тенденцията е те да продължават да нарастват с бързи темпове. Това определя необходимостта от разработването на бързи, паралелни алгоритми за анализ на тези данни. В настоящата публикация е предложен един паралелен алгоритъм, който е модификация на EM-алгоритъма. Решена е класификационна задача от смесен тип – комбинация от машинно самообучение с учител и без учител. Представеният алгоритъм приема на входа данни както с известни, така и с не известни стойности на класификационната функция. В много области класификацията на данни от учител е труден и времеемък процес. Обикновено разполагаме с много на брой примери без дефинирани стойности на класификационната функция и с малко на брой такива, чиито класификации са предварително спицифицирани от учител.*

***Ключови думи**: смесено машинно самообучение, EM-алгоритъм, класификация*

1. Увод

Развитието на уеб технологиите доведе до съществено нарастване на обемите от данни, с които разполагаме. Това доведе до търсене към паралелни алгоритми, които да могат да се справят с наличното количество данни в разумно време. В настоящата публикация е представен точно такъв паралелен алгоритъм в областта на машинното самообучение. Предложена е многонишкова модификация на EM-алгоритъма за смесено машинно самообучение, която обучава класификатор на базата на обучаващо множество от примери, не всички от които имат предварително зададени значения на класификационната функция. Описаната модификация е от изключителна важност, защото бидейки итеративен процес, EM-алгоритъмът изисква голяма изчислителна мощ за намиране на максимално правдоподобна оценка. Поради тази причина, стандартния EM-алгоритъм все още не е масово използван, особено когато става въпрос за големи бази данни. Друго предимство на представения алгоритъм е, че много области от човешкия

живот могат да се възползват от малкото на брой класифицирани примери, достатъчни за постигане добра класификационна точност. Процесът на първоначална класификация на примерите от обучаващото множество може да се окаже дълъг и труден. Необходими са тясно-профилирани знания в конкретната област. Необходим е така наречения „учител“ – който да дава експертно мнение, за да могат да бъдат определени класификациите на примерите. Идеята е необходимостта от такъв учител да се сведе до минимум. Лесно достъпни и “евтини” за използване са примерите без класификации. Социалните мрежи могат да се възползват от връзките между различните потребителите за научаване на дадено понятие. Обикновено потребителите, които са харесали определен продукт са малко или напълно липсват, рейтингът е рядкопопълвана секция от потребителя. Малко хора след покупка се връщат в сайта, от който са пазарували, за да дадат своето мнение и оценка.

Следващата секция прави преглед на изследванията в това направление. Секция 3 представя теоретичната обосновка и дефиниция на поставената задача, описани са и сравнени примери, посредством Наивен Бейсов класификатор (машинно самообучение с учител), паралелен EM-алгоритъм за смесено машинно самообучение. В секция 4 са представени изследвания на класификационната точност на алгоритъма според броя на примерите с класификации в обучаващото множество. Направена е оценка на сложността на алгоритъма. Секция 5 е заключителна и дава предложения за бъдещо развитие.

2. Подобни разработки

EM-алгоритъмът е публикуван през 1977 за първи път от Артур Демпстър [1], но поради своята специфика едва наскоро добива по-голяма популярност и приложимост. P.E. López-de-Teruel, J.M. García и M. Acacio представят парарелен алгоритъм за базовия EM-алгоритъм [12], който разпределя обучаващите примери между паралелни процесори. За разлика от тях, представеният в публикацията модел използва паралелизация по класификационните класове. Xiaojin Zhu прилага графи [10] и хармонична функция [6] за смесено обучение. Също така описва модификация на EM-алгоритъма за смесено машинно самообучение [1], която в настоящата работа е оптимизирана откъм бързодействие.

3. Паралелен EM-алгоритъм за смесен модел на машинно самообучение

Смесеният модел за машинно самообучение получава на входа си два типа данни: примери, класифицирани от учител и такива без предварително зададена класификация.

Определение 1: Пример без предварително зададена класификация: D-измерен вектор $X = (X_1, \ldots, X_d)$, $X \in \mathrm{DM}$ – множеството от допустими стойности на примера.

Няма учител, който да дефинира стойността на класификационната функция. Разполагаме само със стойностите на атрибутите на примера. Ще означаваме такива примери с малко x .

Определение 2: Класифициран от учител пример: (D+1)-измерен вектор $X = (X_1, \ldots, X_d, Y)$, където Y е стойността на дефинирана класификационна функция

В определение 2 разполагаме с учител, който предварително е класифицирал примера X, тоест X принадлежи на класа Y. Казваме още, че значението на класификационната функция на примера X е Y. За краткост ще записваме пример като *(x,y)*, където първата част на наредената двойка съответства на d-измерната част на примера, а y на класификацията на примера.

Тогава под множество от примери ще дефинираме:

$$D = D_1 + D_2$$

където:

Множество от примери от определение 1 – независими и еднакво

$$D_1 = \{(x_i, y_i)\}_{i=1}^{n_l}$$

разпределени, n_l– брой на примерите в това множество

$$D_2 = \{x_j\}_{j=1}^{n_u}$$

Множество от примери от определение 2 – независими и еднакво разпределени, n_u - брой на примерите в това множество

Определение 3: Смесено машинно самообучение (смесен модел от машинно самообучение с учител и без учител) – машинно самообучение, чиито алгоритми използват частичен брой класифицирани примери (подмножество от всичките примери са предварително класифицирани от учител).

3.1. EM-алгоритъм

Итеративна процедура, която намира локален максимум на $\log(X|\theta)$ или $\log(X|\theta) + \log(\theta)$, използвайки метода на максимално правдоподобие. Този алгоритъм има сходимост към локален максимум. Когато примерите са независими и еднакво разпределени:

$$\log P(X \mid \theta) = \log \prod_{i=1}^{n} P(x_i \mid \theta)$$

Стъпки на EM-алгоритъма за смесено машинно самообучение:
Инициализация

$$D = D_1 + D_2 \qquad D_1 = \{(x_i, y_i)\}_{i=1}^{n_l} \qquad D_2 = \{x_j\}_{j=1}^{n_u}$$

Повтори докато $p(D|\theta^t)$ не се променя съществено:

Е-стъпка - намиране на скритото разпределение $p\left(D_2 \middle| D_1, \theta^t\right)$

М-стъпка - намери θ^{t+1}, така че

$$\sum_{D_2} p\left(D_2 \middle| D_1, \theta^t\right) \log\left(pD_1, D_2 \middle| \theta^{t+1}\right) \text{ е максимално}$$

В проведената експериментална база е използван смесен модел от нормални разпределения. Тогава параметърът $\theta = (\pi, \mu, \Sigma)$, където

π – априорна вероятност за класа

μ – математическо очакване

Σ – ковариационна матрица

$$N(x, \mu_y, \Sigma_y) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\Sigma_y\right|^{\frac{1}{2}}} e^{\frac{-(x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y)}{2}}$$

3.2. Паралелен EM-алгоритъм за смесено машинно самообучение (Multi-threaded Semi-supervised Expectation Maximization Algorithm - MT-SSEM)

Псевдокод на предложения паралелен EM алгоритъм за смесено машинно самообучение:

1. ParalelInit() - Инициализация (използва само частта от примери с известни класификации). Паралелизация по класовете j.

$$\pi_j = \frac{\text{броят примери от клас } j}{\text{общият брой примери}}$$

$$\mu_j = \frac{\sum x_{ij}}{\text{броят примери от клас } j}$$

$$\Sigma_j = \frac{\sum (x_{ij} - \mu_j)(x_{ij} - \mu_j)^t}{\text{броят примери от клас } j}$$

2. Докато няма промяна в $p\left(D|\theta^t\right)$:
   - *Parallel Expectation();* - намери паралелно $P\left(y_j \mid x_i, \theta\right)$
   - *Parallel Maximization();* - намери паралелно θ_j

На всяка стъпка на алгоритъма се преизчислява първо паралелно $P\left(y_j \mid x_i, \theta\right)$, а после намира параметъра θ_j за всеки един клас едновременно.

Представеният по-горе алгоритъм е лесно приложим и интегрируем в стандартна компютърна архитектура с многоядрен процесор. Реализацията на MT-SSEM представлява библиотека, отделен многонишков модул, написан на C++. Използвани са стандартната C++ библиотека - stl, eigen – включва голяма база от функционалност, засягаща линейната алгебра – [8].

Тестовете са проведени върху компютърна архитектура, състояща се от процесор i7-2630QM с 4 ядра (8 нишки), 8GB RAM. За оптималност е използвана база данни с примери, разпределени в 8 класа (броя нишки на процесора).

4. Проведени експерименти и резултати

4.1. Класификационна точност

Поведение и изследване на алгоритъма върху ръчно генерирано тестово множество с атрибути, които се подчиняват на нормално разпределение:

Тестовото множество се състои от 5000 примера, разпределени в 8 класа. Броят на атрибутите е 5, стойностите на всеки от тях са получени с генератор на нормално разпределение. Използвана е линейна регресия, след това е приложен процес на дискретизация в 8 класа. В таблицата по-долу се вижда изменението на класификационната точност на алгоритъма върху тестовото множество.

Таблица 1. Класификационна точност на алгоритъма според процентния брой използвани класифицирани примери

Labeled %	MT-SSEM %
10%	92.28%
25%	92.84%
33%	94.63%
50%	95.95%
75%	97.23%
100%	98.48%

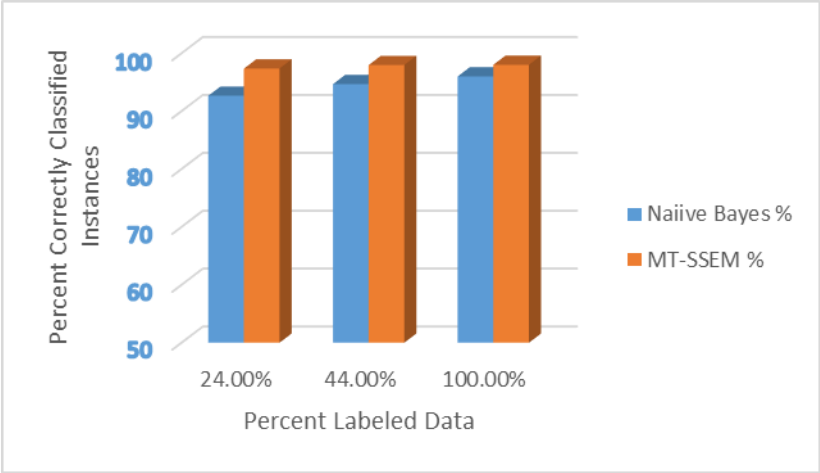
Поведение и изследване на алгоритъма върху тестово множество, предоставено от университета Waikato – “iris.arff” [5]*:*

В графиките по-долу MT-SSEM е сравнен с представянето на Наивен Бейсов модел. С намаляване на броя класифицирани примери, тенденцията е към по-стръмно падане на класификационната точнност на наивния подход за разлика от EM-алгоритъма за смесено машинно самообучение, който се представя доста стабилно.

Проведени са 3 опита – както следва:

- 100% от обучаващото множество се ползва – всички примери имат класификации. Тогава паралелният EM-алгоритъм за смесено машинно самообучение – MT-SSEM показва 98% класификационна точност върху обучаващото множество. Резултатите са без прилагане на процедура по крос валидация. Наивният Бейсов модел се представя също добре (използвана е реализация на weka – графичен интерфейс) – 96% точност
- 44% от обучаващото множество се ползва – класификационната точност на Наивния модел пада значително за разлика от тази на MT-SSEM.
- 24% от обучаващото множество се ползва – запазва се тенденцията към по-стръмно падане на наивния подход.

Фигура 1. Сравнение между Наивен Бейсов Модел и MT-SSEM. По оста x са процента примери без определени предварително класификации, а по оста y е



класификационната точност върху тестовото множество.

Таблица 2. Класификационна точност на Наивен Бейсов Класификатор и MT-SSEM в проценти – брой правилно класифицирани примери в зависимост от процентното съдържание на броя класифицирани примери

Labeled examples %	Naiive Bayes %	MT-SSEM %
24.00%	92.66	97.35
44.00%	94.67	97.96
100.00%	96.00	98.00

4. 2. Паралелизъм

Резултатите са само върху фазата на учене. Фазата на класификация - определянето на класификацията на пример след проведено обучение е бързо, със сложност O(|Y|), където |Y| е броят класове в множеството. Не е необходима паралелизация.

Сложността на EM алгоритъма за смесено машинно самообучение:

- *Parallel Expectation():* Стандартният алгоритъм има сложност $O(|Y|*n_u)$, където n_u е броят примери с неизвестни класификации. Благодарение на паралелизма сложността се свежда до $O(n_u)$

- *Parallel Maximization():* Стандартният алгоритъм има сложност O(|D|*|Y|): където |D| е общият брой примери в обучаващото множество, |Y| е броят различни класове. С помощта на паралелната версия тази сложност се намалява до O(|D|).

5. Заключение

EM-алгоритъмът и неговите модификации, в частност MT-SSEM все още не са широко застъпени в реалните системи, използват се по-евтини откъм изчислителна мощ алгоритми. С навлизането на технологиите и най-вече hadoop, се отвориха портите на една настояща мода към паралелизъм и разпределение на работата на алгоритмите. Тези системи са модерни и приложими, но изискват познания по поддръжка, инсталиране и използване на средата. За разлика от тях, предложеният алгоритъм в настоящата работа може да се интегрира лесно към всяка работна станция и показва много добра класификационна точност. Приложим е при решаване на класификационни задачи, които имат ограничен брой класифицирани примери. Бъдещето развитие и предизвикателство е реално приложение на всички модификации на EM в много големи бази данни от естеството на тези, с които разполагат google, amazon и др, при които фазата на обучение трябва да се изпълнява доста често.

Благодарности

Изследването е осъществено с помощта на Европейския социален фонд чрез Оперативна програма „Развитие на човешките ресурси”, договор № BG051PO001-3.3.06 - 0052 (2012-2014).

Литература

1. Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009
2. Olivier Chapelle, Bernhard Scholkopf, Alexander Zien "Semi-supervised Learning", 2006, MIT Press
3. Ian Witten, Eibe Frank, Practical Machine Learning Tools and Techniques, Morgan Kaufmann. 2011. (Third Edition)
4. Димитър Вандев, „Записки по приложна статистика 1“, 2003
   http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/lectures/applstat1.pdf
5. Weka - http://www.cs.waikato.ac.nz/ml/weka/
6. Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *The 20th International Conference on Machine Learning (ICML)*, 2003.
7. Multivariate normal distribution

8. http://en.wikipedia.org/wiki/Multivariate_normal_distribution
9. Библиотека за линейна алгебра - http://eigen.tuxfamily.org/index.php?title=Main_Page
A. P. Dempster, N. M. Laird and D. B. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, B, vol. 39, no. 1, pp. 1–38.
10. Xiaojin Zhu. Semi-Supervised Learning with Graphs. PhD thesis, Carnegie Mellon University, 2005. CMU-LTI-05-192.
11. Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, 2003.
12. P. L´opez-de-Teruel, J. Garc´ıa, and M. Acacio. The parallel EM algorithm and its applications in computer vision. *Proceedings* of the PDPTA'99, CSREA Press, 1999.

PARALLEL SEMI-SUPERVISED EM-ALGORITHM (MT-SSEM)

Gergana Lazarova[1], Ivan Koychev[1,2]

[1]Sofia University "Kliment Ohridski"
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
gerganal@fmi.uni-sofia.bg, koychev@fmi.uni-sofia.bg

***Abstract**: With the development of new technology, information has become easier to access. Databases consist of numerous examples, reaching new dimensions. New powerful, parallel, executing fast algorithms are needed. The aim of the presented algorithm is exactly to fulfill the greed for this trend. A multi-threaded semi-supervised version of the standard EM-algorithm is presented. It uses two data sources – labeled and unlabeled data. Usually, we have access to lots of unlabeled examples, but the labeled ones are difficult to reach.*

***Keywords**: semi-supervised machine learning, EM-algorithm, classification*