様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３０　日現在

機関番号：　１２６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０１３２５

研究課題名（和文）オンライン学習および転移学習の併用による画像診断支援システムの動的高性能化

研究課題名（英文）Dynamic performance improvement in computer-assisted detection system for diagnostic imaging by combined application of online learning and transfer learning

研究代表者

野村　行弘（Nomura, Yukihiro）

東京大学・医学部附属病院・特任研究員

研究者番号：６０４３６４９１

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究は，画像診断装置や撮像条件の違いによる画像データの多様性に対応したコンピュータ支援検出(CAD)ソフトウェアの動的高性能化の方法論について検討した．オンライン転移学習アルゴリズムを用いたCADソフトウェアの学習方法を構築し，多施設データによる検証を行った．また，画素単位で学習を行う識別器を用いたCADソフトウェアの性能改善に必要となる病変形状情報入力を省力化した場合の病変検出性能への影響について検討した．

研究成果の概要（英文）：In this research, a dynamic performance improvement of computer-assisted (CAD) detection software by considering diversity of image data due to imaging scanner or imaging parameters was investigated. We constructed a training method of CAD software using online transfer learning algorithm, and verified by multicenter data. We also investigated whether a simplified method of gold standards definition can be used as an alternative to gold standards defined by pixel-by-pixel painting for CAD software using voxel-based classification.

研究分野：　医用画像処理

キーワード：　医用画像　診断支援システム　転移学習　オンライン学習

１．研究開始当初の背景

医用画像におけるコンピュータ支援検出（computer-assisted detection, CAD）ソフトウェアの開発は画像診断の効率・精度の向上を期待して進められており，大きな臨床的意義があると考えられている．CAD ソフトウェアの性能は機械学習によるところが大きく，CAD ソフトウェア開発用データセットと実運用時のデータの画質が異なる場合，期待される性能が得られないことがある．

研究代表者らはこれまでに，多施設臨床使用で収集した CAD ソフトウェアの処理結果および医師の診断結果に基づく評価データを用いた識別器の再学習を行うことで施設毎の CAD ソフトウェアの性能改善を図った．画像診断装置や撮像技術は年々進歩している．このため，識別器の再学習による CAD ソフトウェアの性能改善を継続して行うことが不可欠であるが，以下の点が課題として挙げられた．

(a)　現状は各施設のデータを研究代表者の所属施設に集約した上で，一括で再学習を行っている．このため，施設数が増えると集約した学習データの管理が問題となる．特に，画素単位で学習を行う識別器の場合，データは膨大なものとなる．従って，各施設で新たに発生したデータを施設内で逐次的に学習することが望ましい．このためには，逐次的な学習を行うための環境の構築が必要である．

(b)　従来の新規施設のデータを用いた CAD ソフトウェアの性能改善では，試行錯誤により選択した学習データセットを用いた再学習を行ってきた(図 1(a))．機械学習技術には，ある問題を効果的かつ，効率的に解くために，別の関連した問題の学習結果を利用する転移学習が長く研究されており，新規施設に向けた CAD ソフトウェアの性能改善に転移学習を適用すれば，効率良く学習が行えると考えられる(図 1(b))．

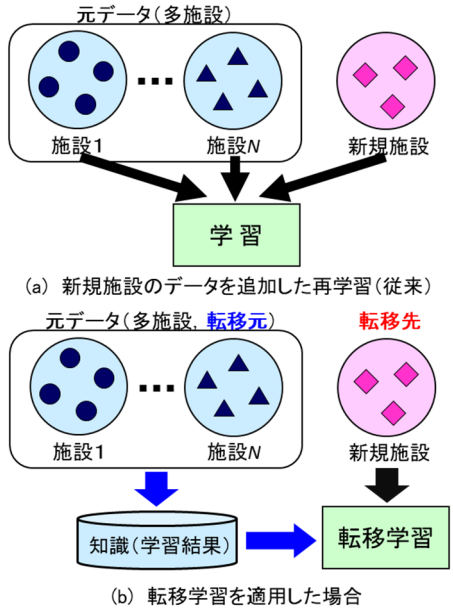

図 1 新規施設のデータを用いた CAD ソフトウェアの性能改善

２．研究の目的

本研究の目的は，画像診断装置や撮像条件の違いによる画像データの多様性に対応した CAD ソフトウェアの動的高性能化を図るための方法論の構築である．実現に向けて主に以下の項目について研究を実施した．

(1) オンライン転移学習アルゴリズムを用いた学習方法の構築
(2) 病変形状情報入力の省力化による病変検出性能への影響についての検討

３．研究の方法

(1)　オンライン転移学習アルゴリズムを用いた学習方法の構築

図 2 に構築したオンライン転移学習アルゴリズムの擬似コードを示す．転移元および転移先（新規施設）において，それぞれのデータのみで学習した識別器に関する情報は，*src* および *tgt* を上付き添え字として表記し，転移学習に関しては *trs* と表記する．例えば，転移元，転移先のそれぞれのデータのみで学習した識別器はh^{src}，h^{tgt}と書く．各識別器の出力は$s(x) = \max\{1, \min\{0, h(x)\}\}$によって$[0, 1]$のスコア値に変換する．

症例単位の学習を実現するために ROC 曲線の下面積(AUC)ベースの損失関数を適用する．t 例目のデータが m 個の真陽性(TP)データ$\{x_i^+\}_{i=1}^m$と n 個の偽陽性(FP)データ$\{x_k^-\}_{k=1}^n$で構成されるとき，損失関数は以下の式で定義される．

$$\ell_t^{alg} = \frac{1}{mn}\sum_i^m \sum_k^n \max\left(0,\ s_t^{alg}(x_k^-) - s_t^{alg}(x_i^+)\right) \quad (1)$$

ただし，$alg \in \{src, tgt\}$である．識別器出力は各識別器の重み付き平均となる．

$$s_t^{trs}(x) = \frac{w_t^{src}}{w_t^{src} + w_t^{tgt}} h^{src}(x) + \frac{w_t^{tgt}}{w_t^{src}+w_t^{tgt}} h_t^{tgt}(x) \quad (2)$$

構築したアルゴリズムは負の転移に関する理論保証が得られているが，紙面の都合上省略する．

1: Set $w_0^{src} = 1, w_0^{tgt} = 0$.
2: Set $\rho = 1/(T-1)$, and $\eta = \sqrt{-\frac{8}{T}\log(\rho(1-\rho)^{T-2})}$.
3: **for** each round $t = 1, \ldots, T$ **do**
4: $\quad w_t^{src} = (1-\rho) w_{t-1}^{src} \exp(-\eta \ell_t^{src}) + \rho w_{t-1}^{tgt} \exp(-\eta \ell_t^{tgt})$,
5: $\quad w_t^{tgt} = (1-\rho) w_{t-1}^{tgt} \exp(-\eta \ell_t^{tgt}) + \rho w_{t-1}^{src} \exp(-\eta \ell_t^{src})$.
6: **end for**

図 2　構築アルゴリズムの擬似コード

構築したアルゴリズムについて，頭部 MRA 画像の脳動脈瘤検出ソフトウェアの FP 削減処理用識別器（識別器 AdaBoost，63 特徴量）を対象としたシミュレーション実験を行った．症例データは表 1 に示す 4 施設より収集したデータ（CAD 処理結果および読影医が入力したフィードバックデータ）を使用した．AdaBoost の弱識別器数は転移元，転移先ともに 100 とし，提案手法のパラメータ T は 100

とした．学習データは FP データを TP データと同数となるようにランダムサンプリングにより削減した．

転移元（東大病院）では，学習データ 400 例を用いて転移元識別器h^{src}の学習を行った．なお，学習データのランダムサンプリングを 100 回試行し，評価症例 100 例で性能が最良となった識別器を用いた．

転移先（施設 A～C）ではまず，学習データを 100 例より重複無しで 50 例選択した．選択した症例のうち最初の 5 例を用いてh_t^{tgt}の初期学習を行った．その後，1 例追加する毎に重みw_t^{src}, w_t^{tgt}の更新，および転移先識別器h_t^{tgt}の学習を行い，評価用症例 100 例で評価を行った．なお，学習データの選択および順番はランダムとし，100 回試行した．

評価指標として ROC 曲線の下面積(AUC)を用いた．転移元，転移先の両データを用いた再学習，および Zhao らのオンライン転移学習を比較対象とした．なお，Zhao らのオンライン転移学習の損失関数は提案手法と同一とした．

CT の被ばく低減技術として近年注目されている，逐次近似画像再構成法を用いた CT 画像に対して肺結節検出ソフトウェアを適用した場合の性能評価を行った．複数の再構成条件，X 線量の組合せについて評価した結果，条件によっては画質の変化による性能低下が認められ，CAD ソフトウェアの再学習または転移学習による性能改善の必要性が示唆された．しかし，シミュレーション実験に必要な症例数が収集できなかったため，検討から外した．

表 1　施設間の装置，撮像条件の違い

施設名	装置 vendor	静磁場強度[T]	画素サイズ	
			x,y [mm]	z[mm]
東大病院	GE	3	0.469	0.8
施設A	GE	3	0.391-0.41	0.5
施設B	東芝	1.5	0.367-0.371	0.5-0.55
施設C	Philips	1.5	0.5-0.55	0.5

(2) 病変形状情報入力の省力化による病変検出性能への影響についての検討

画素単位で学習を行う識別器を用いた CAD ソフトウェアの学習には画素単位の病変形状の情報が必要となる．理想的には医師がペイント入力した病変形状を用いることが望ましい．しかし，医師が労力の大きいペイント入力を多数症例に行うことは困難であるため，病変形状入力の省力化が必要である．最も省力化した病変形状の定義方法として，病変の中心点とサイズ情報から生成した球状領域が挙げられる（図 3）．そこで，球状領域で定義した病変形状情報がペイント入力の代替となり得るかを検討した．

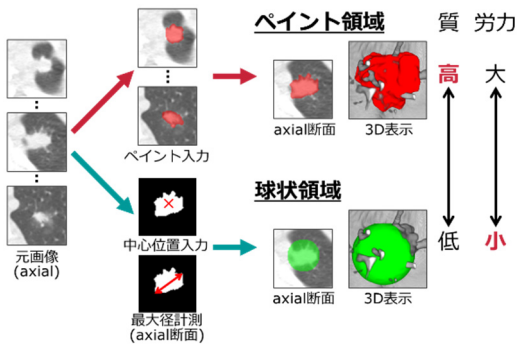

図 3　病変形状情報入力におけるペイント入力と球状領域との違い

対象は画素単位で学習を行う識別器を用いた 2 種類の CAD ソフトウェア（頭部 MRA 画像の脳動脈瘤検出，胸部 CT 画像の肺結節検出）とし，病変形状がペイント入力されているデータセットを使用した．球状領域を定義するための位置およびサイズを，ペイント入力された病変形状の axial 断面での面積が最大となるスライスの重心および最大径を用いた．

実験はまず，各データセットより学習症例を重複無しで 300 例選択し，残りを評価症例とした．選択した症例のうち最初の 50 症例を用いて初期学習を行い，評価用症例で評価を行った．その後，学習症例を 50 症例を追加する毎に同様の学習と評価を繰り返し行った．なお，学習データの選択および順番はランダムとし，50 回試行した．

４．研究成果

(1) オンライン転移学習アルゴリズムを用いた学習方法の構築

図 4 に施設 B を転移先ドメインとした転移学習結果を示す．提案アルゴリズムによるオンライン転移学習では，転移元，転移先の両データを用いた再学習と同等以上の性能が得られた．Zhao らのオンライン転移学習では学習初期に負の転移と呼ばれる，転移元データのみで学習した識別器より性能が低下する現象が発生しているが，構築したアルゴリズムによるオンライン転移学習では負の転移は抑制された．転移先が他の施設の場合も同様の結果が得られた．

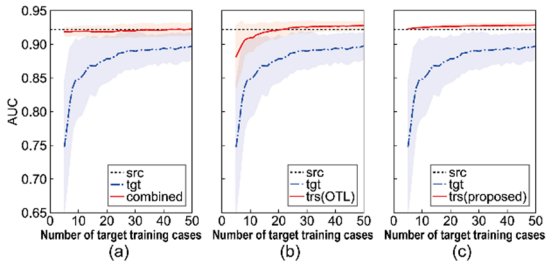

図 4　学習曲線（転移先：施設 B）(a) 転移元･転移先両データを用いた再学習，(b) Zhao らのオンライン転移学習，(c) 構築したアルゴリズムによる転移学習

(2) 病変形状情報入力の省力化による病変検出性能への影響についての検討

図 5 に脳動脈瘤検出の場合の学習曲線を示す．ANODE スコアは[0,1]の範囲で高いほど性能が良いことを示す．この結果より，ペイント入力した 50 症例で得られる性能を球状領域で得るためには約 100 症例必要であること

が示された．図 6 は肺結節検出の場合で，ペイント入力した 50 症例で得られる性能を球状領域で得るためには約 200 症例必要であることが示された．2 種類の CAD で必要症例数に差が出た一因としては，脳動脈瘤の形状が球状に近いのに対して，肺結節の形状は脳動脈瘤より複雑であることが挙げられる．

以上の結果から，病変形状が球状に近い，もしくは症例数が多い場合は球状領域による病変形状がペイント領域の代替となり得ることが示された．

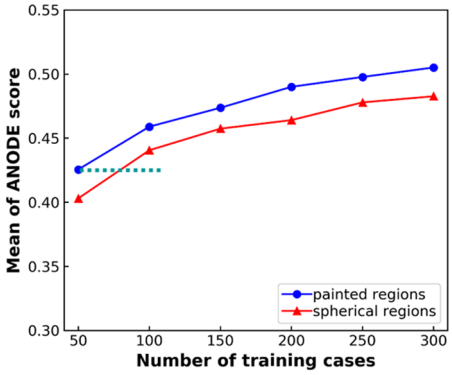

図 5　学習曲線(脳動脈瘤検出)

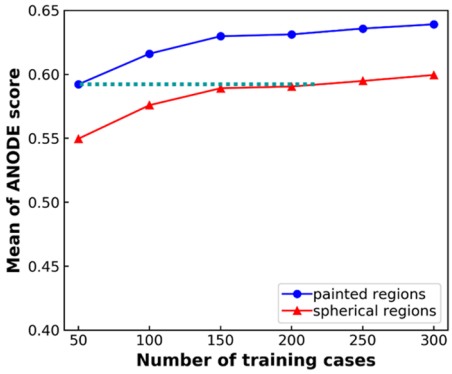

図 6　学習曲線(肺結節検出)

５．主な発表論文等

〔雑誌論文〕(計２件)

① 吉川健啓、野村行弘、佐藤一誠、林直人、花岡昇平、三木聡一郎、阿部修、オンライン学習および転移学習の併用による画像診断支援システムの動的高性能化、査読無、臨床放射線、Vol.62、1237-1243、2017.

② Nomura Y, Higaki T, Fujita M, Miki S, et al., Effects of iterative reconstruction algorithms on computer-assisted detection (CAD) software for lung nodules in ultra-low-dose CT for lung cancer screening, 査読有, Academic Radiology, Vol. 24, No.2, 124-130, 2017.

〔学会発表〕(計５件)

① Sato I, Nomura Y, Hanaoka S, Miki S, et al. Managing computer-assisted detection system based on transfer learning with negative transfer inhibition, KDD 2018, 2018 (採択決定)

② 野村行弘、佐藤一誠、花岡昇平、三木聡一郎　他、オンライン転移学習を用いた脳動脈瘤検出ソフトウェアの性能改善、電子情報通信学会医用画像研究会、2017.

③ Nomura Y, Hayashi N, Hanaoka S, Nemoto M, Takenaga T, Miki S, et al., Effects of different types of gold standard on computer-assisted detection of lung nodules using voxel-based classification, CARS 2017、2017.

④ 佐藤一誠, 野村行弘, 林直人, オンライン転移学習と医用画像読影支援への応用、日本応用数理学会 2016 年度年会, 2016.

⑤ Nomura Y, Miki S, Sato I, et al. Building an integrated CAD development environment in clinical settings (the 9th report): new web-based image database system for CAD development, 第 75 回日本医学放射線学会総会, 2016.

６．研究組織

(1)研究代表者

野村　行弘 (NOMURA, Yukihiro)
東京大学・医学部附属病院・特任研究員
研究者番号：６０４３６４９１

(2)連携研究者

佐藤　一誠 (SATO, Issei)
東京大学・大学院新領域創成科学研究科・講師
研究者番号：９０６１０１５５

三木　聡一郎 (MIKI, Soichiro)
東京大学・医学部附属病院・特任助教
研究者番号：３０７０７７６６