Universitatea Politehnica București
Facultatea de Automatică și Calculatoare
Departamentul de Automatică și Ingineria Sistemelor

Rezumatul Tezei de Doctorat

Strategii Adaptive pentru Antrenarea Dicționarelor Separabile și Aplicații pentru Detecția de Anomalii

Andra-Elena Băltoiu

Coordonator:

Prof. dr. ing. Bogdan Dumitrescu

2021

1 Introducere

Antrenarea dicționarelor pentru reprezentări rare (DL) reprezintă o clasă de metode de prelucrare de semnale ce prespun aproximarea acestora utilizând o combinație liniară de câteva elemente ale unei baze, denumită dicționar. Problema de antrenare prespune ca atât dicționarul, cât și reprezentarea rară să fie învățate în baza datelor. Modelul este folosit pentru a rezolva sarcini generale de reconstrucție de semnale, în probleme de clasificare, aplicații ale prelucrării imaginilor și în compressed sensing. Însă semnalele în două sau mai multe dimensiuni necesită vectorizare prealabilă pentru a putea fi utilizate în modelul standard DL. Această operație poate distruge corelațiile prezente în date în cea de-a doua dimensiune. O soluție la acest inconvenient o reprezintă modelul dicționarelor separabile (SDL), în care dicționarul este structurat, luând forma unui produs Kronecker de două dicționare cu dimensiune mai mică, ceea ce conduce și la un avantaj computațional.

Modelul SBL reprezintă tema centrală a tezei. În particular, contribuțiile se referă la determinarea valorilor optimele pentru cei doi parametrii: sparsitatea reprezentării și dimensiunea dicționarului. Principala ipoteză a modelului DL o reprezintă existența unei reprezentări rare a semnalelor, însă aceasta poate fi demonstrată în mod riguros doar pentru anumite clase de semnale. Experimentele practice au arătat utilitatea modelului și în cazul multor alte aplicații. Problema alegerii valorii sparsității rămâne deci deschisă, posibilitățile de a determina precis valoarea acesteia fiind limitate. Cel de-al doilea parametru îl reprezintă numărul de elemente al dicționarului (numite atomi). Testele numerice [1] arată că deși dicționarele de dimensiuni mari produc aproximări mai bune, îmbunătățirile se plafonează după o anumită valoare a dimensiunii, încât efortul computațional suplimentar devine nejustificat.

O altă contribuție a lucrării o reprezintă adaptarea modelului separabil la problema identificării anomaliilor pe grafuri. Aceasta presupune identificarea topologiilor anormale, anume structuri de (sub-)graf ce se deosebesc considerabil de restul tiparelor de conectivitate. În acest caz, semnalele de tip graf pot fi reprezentate utilizând matricile Laplacian, astfel încât atomii dicționarului rezultat să descrie configurații elementare de legături între noduri.

Problema detecției de anomalii este abordată în lucrare în alte două contribuții, ce iau în considerare cerințele curente pentru metode online și nesupervizate.

Problema antrenării dicționarelor se poate formaliza

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X} + \boldsymbol{V}. \tag{1}$$

unde $\boldsymbol{Y} \in \mathbb{R}^{m \times N}$ reprezintă cele N semnale, $\boldsymbol{D} \in \mathbb{R}^{m \times n}$ este dicționarul, având n atomi, $\boldsymbol{X} \in \mathbb{R}^{n \times N}$ este matricea de reprezentări rare, iar $\boldsymbol{V} \in \mathbb{R}^{m \times N}$ este zgomotul, de cele mai multe ori considerat a fi Gaussian.

Atât dicționarul, cât și reprezentarea sunt învățate, ceea ce conduce la următoarea problemă de optimizare

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{X}} \ & \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2 \\ s.t. \ & \|\boldsymbol{x}_l\|_0 \leq s, l = 1:N \\ & \|\boldsymbol{d}_j\|_2 = 1. \end{aligned} \tag{2}$$

Cu toate că $\boldsymbol{D}$ poate fi construit în prealabil, tratăm exclusiv cazul în care acesta este antrenat, această abordare având avantajul unor aproximări mai exacte. Problema de mai sus este de regulă rezolvată folosind o schemă de minimizare alternativă, în care $\boldsymbol{D}$ și $\boldsymbol{X}$ sunt modificate în mod iterativ până la îndeplinarea unei condiții. Procedura începe cu inițializarea dicționarului cu elemente aleatoare și calculul reprezentării. La iterația următoare, forma lui obținută anterior $\boldsymbol{X}$ se fixează, iar dicționarul este actualizat. În mod curent, condiția de oprire este dată de numărul de iterații, având în vedere că majoritatea algoritmilor pentru actualizarea lui $\boldsymbol{D}$ converg către soluție după un număr suficient de iterații.

În modelul separabil, dicționarul are forma $\boldsymbol{D} = \boldsymbol{D}_2 \otimes \boldsymbol{D}_1$, cu $\boldsymbol{D}_1 \in \mathbb{R}^{m_1 \times n_1}$ și $\boldsymbol{D}_2 \in \mathbb{R}^{m_2 \times n_2}$. Problema devine

$$\boldsymbol{Y} = \boldsymbol{D}_1 \boldsymbol{X} \boldsymbol{D}_2^\top + \boldsymbol{V} \tag{3}$$

și este echivalentă cu (1), având în vedere proprietățile produsului Kronecker

$$\text{vec}(\boldsymbol{D}_1 \boldsymbol{X} \boldsymbol{D}_2^\top) = (\boldsymbol{D}_2 \otimes \boldsymbol{D}_1)\text{vec}(\boldsymbol{X}). \tag{4}$$

2 Sparsity Bayesian Learning pentru Dicționare Separabile

Prima constrângere din (2) presupune că nivelul de sparsitate s este cunoscut, ceea ce se întâmplă rar în aplicațiile practice. Însă performanțele algoritmilor sunt influențate de alegerea lui s, în special în anumite clase de probleme, cum sunt aplicațiile de compressed sensing. Prezentăm un experiment care arată că estimarea eronată a sparsității duce fie la overfitting, fie la underfitting. Un set de semnale sintetice este construit în baza unui nivel de sparsitate cunoscut, iar experimentul evaluează diferite modele (în care s ia diferite valori) utilizând raportul dintre eroare de antrenare și cea de test. Rezultatele arată că subestimarea nivelului sparsității duce la valori similare pentru cele două erori, semnalând faptul că modelul nu este antrenat în mod adecvat, în timp ce supraestimarea lui s duce la overfitting.

Soluțiile existente pentru rezolvarea acestor probleme presupun fie estimarea lui s într-o etapă preliminară, utilizând un set restrâns de semnale pentru a stabili limitele sparsității, fie utilizează strategii adaptive ce estimează sparsitatea optimă pe măsură ce antrenarea avansează. O a treia abordare presupune interpretarea lui (2) ca o problemă de Sparse Bayesian Learning (SBL) care, în formularea ierarhică, conduce la o soluție rară fără a necesita utilizarea unei

probabilități a priori care să inducă sparsitate. Rescriem problema DL pentru cazul unui singur semnal $\boldsymbol{y}$, în locul matricii de semnale $\boldsymbol{Y}$, și reprezentarea corespunzătoare a acestuia, $\boldsymbol{x}$.

Abordarea standard SBL (non-ierarhică), presupune găsirea lui $\boldsymbol{x}$ astfel

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}). \tag{5}$$

Această formulare conduce la soluția Maximum a Posteriori (MAP), ce poate fi însă îmbunătățită, dacă se introduce un hiperparametru $\boldsymbol{\gamma}$ ce controlează varianța fiecărui element al reprezentării

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\boldsymbol{y}, \sigma^2) = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{y}|\boldsymbol{\gamma}, \sigma^2)p(\boldsymbol{\gamma}). \tag{6}$$

Este demonstrat în [2] că o probabilitate a priori neinformativă $p(\boldsymbol{\gamma})$ duce la o soluție rară, astfel că nu este necesară cunoașterea valorii reale de sparsitate. Soluția optimă $\boldsymbol{\gamma}$ presupune estimarea probabilității (Gaussiene) a posteriori $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\gamma}, \sigma^2)$, de medie $\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}_x\boldsymbol{D}^\top\boldsymbol{y}$ și varianță

$$\boldsymbol{\Sigma}_x = \left(\sigma^2\boldsymbol{D}^\top\boldsymbol{D} + \boldsymbol{\Gamma}^{-1}\right)^{-1}. \tag{7}$$

unde $\boldsymbol{\Gamma}$ este o matrice diagonală conținând valorile hiperparametrilor. Calculul matricii de covarianță de mai sus este costisitor din punct de vedere computațional, datorită dimensiunii acesteia. O formulare alternativă, utilizând covarianța $\boldsymbol{\Sigma}_y$, de dimensiuni mai mici, este propusă în [3].

$$\boldsymbol{\Sigma}_y = \sigma^2\boldsymbol{I} + \boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}^\top. \tag{8}$$

Expectanța devine

$$\boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{D}^\top\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}. \tag{9}$$

Pentru estimarea lui $\boldsymbol{\gamma}$ se pot utiliza diferite abordări [2, 3]. În cele ce urmează vom utiliza algoritmul Expectation-Maximization (EM) (k desemnează numărul iterației curente):

$$\text{E step: } \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\gamma}^{(k)}}[x_i^2] = (\Sigma_x)_{i,i} + \mu_i^2, \tag{10}$$

$$\text{M step: } \gamma_i^{(k+1)} = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\gamma}^{(k)}}[x_i^2], \tag{11}$$

$$(\sigma^2)^{(k+1)} = \frac{1}{m}\left(\|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\mu}\|^2 + (\sigma^2)^{(k)}\sum_{i=1}^{n}\left(1 - \frac{\Sigma_{x_{i,i}}}{\gamma_i}\right)\right)^{1/2}. \tag{12}$$

Contribuția noastră (SBL-2D) [4] constă în adaptarea metodei SBL pentru cazul dicționarelor separabile și presupune două etape. Pentru început, suportul reprezentării este identificat în mod eficient din punct de vedere computațional, prin modificarea modului în care hiperparametrii controlează elementele reprezentării. Datorită faptului că modelul implică două dicționare, fiecare element din $\boldsymbol{X}$ este influențat de un atom din $\boldsymbol{D}_1$ (mai exact o coloană a dicționarului) și

un atom din $\boldsymbol{D}_2$ (o linie din $\boldsymbol{D}_2^\top$). Propunem ca varianța liniilor și coloanelor din $\boldsymbol{X}$ să fie controlată în mod independent de doi hiperparametri, $\boldsymbol{\beta}^{(1)} \in \mathbb{R}^{n_1}$ pentru linii și $\boldsymbol{\beta}^{(2)} \in \mathbb{R}^{n_2}$ pentru coloane.

Pentru a adapta relațiile (7-8), matricea $\boldsymbol{\Gamma}$ este înlocuită cu $\boldsymbol{B}^{(d)}$, câte o matrice pentru fiecare proces SBL. Fiecare $\boldsymbol{B}^{(d)}$ poate fi exprimată ca produs Kronecker, notat general $\boldsymbol{B}^{(d)} = \boldsymbol{B}_2^{(d)} \otimes \boldsymbol{B}_1^{(d)}$ și având formele particulare

$$\begin{aligned} \boldsymbol{B}^{(1)} &= \boldsymbol{I}_{n_2} \otimes \operatorname{diag}(\boldsymbol{\beta}^{(1)}), \\ \boldsymbol{B}^{(2)} &= \operatorname{diag}(\boldsymbol{\beta}^{(2)}) \otimes \boldsymbol{I}_{n_1}. \end{aligned} \tag{13}$$

Notația $\boldsymbol{I}_{n_d}$ desemnează matricea identitate de dimensiune n_d. Prima relație de mai sus indică faptul că elementele liniei i din $\boldsymbol{X}$ sunt asociate cu $\beta_i^{(1)}$. Introducem notația $\boldsymbol{\beta}^{(d)}$, $d \in \{1, 2\}$ pentru a face referire la ambele dimensiuni și o extindem și în cazul celorlalte variabile ce conțin valori atât pentru linii cât și pentru coloane.

Utilizând proprietăți cunoscute ale produsului Kronecker, adaptarea (8) pentru cazul 2D devine

$$\boldsymbol{\Sigma}_y^{(d)} = (\sigma^{(d)})^2 \boldsymbol{I} + (\boldsymbol{D}_2 \boldsymbol{B}_2^{(d)} \boldsymbol{D}_2^\top) \otimes (\boldsymbol{D}_1 \boldsymbol{B}_1^{(d)} \boldsymbol{D}_1^\top). \tag{14}$$

Matricile de covarianță structurate sunt comune în multe probleme multivariabile. Prin urmare, calculul inversei unei astfel de matrici, cum este (14), se poate efectua în mod eficient utilizând descompunerea valorilor singulare (SVD) [5]

$$\begin{aligned} \boldsymbol{D}_1 \boldsymbol{B}_1^{(d)} \boldsymbol{D}_1^\top &= \boldsymbol{U}_1^{(d)} \boldsymbol{S}_1^{(d)} \boldsymbol{U}_1^{(d)\top}, \\ \boldsymbol{D}_2 \boldsymbol{B}_2^{(d)} \boldsymbol{D}_2^\top &= \boldsymbol{U}_2^{(d)} \boldsymbol{S}_2^{(d)} \boldsymbol{U}_2^{(d)\top}. \end{aligned} \tag{15}$$

Utilizând proprietatea de ortogonalitate a matricilor de mai sus, inversarea matricii de covarianță devine

$$\left(\boldsymbol{\Sigma}_y^{(d)}\right)^{-1} = (\boldsymbol{U}_2^{(d)} \otimes \boldsymbol{U}_1^{(d)}) \left((\sigma^{(d)})^2 \boldsymbol{I} + \boldsymbol{S}_2^{(d)} \otimes \boldsymbol{S}_1^{(d)}\right)^{-1} (\boldsymbol{U}_2^{(d)} \otimes \boldsymbol{U}_1^{(d)})^\top, \tag{16}$$

astfel că singura inversare necesară este cea a matricii diagonale, pe care o notăm

$$\left((\sigma^{(d)})^2 \boldsymbol{I} + \boldsymbol{S}_2^{(d)} \otimes \boldsymbol{S}_1^{(d)}\right)^{-1} = \operatorname{diag}(\boldsymbol{t}^{(d)}),$$

cu $\boldsymbol{t}^{(d)} \in \mathbb{R}^m$. Folosind expresia de mai sus și notând

$$\boldsymbol{E}_1^{(d)} = \boldsymbol{B}_1^{(d)} \boldsymbol{D}_1^\top \boldsymbol{U}_1^{(d)}, \quad \boldsymbol{E}_2^{(d)} = \boldsymbol{B}_2^{(d)} \boldsymbol{D}_2^\top \boldsymbol{U}_2^{(d)}, \tag{17}$$

soluția (9) devine

$$\boldsymbol{\mu}^{(d)} = (\boldsymbol{E}_2^{(d)} \otimes \boldsymbol{E}_1^{(d)}) \cdot \operatorname{diag}(\boldsymbol{t}^{(d)}) \cdot (\boldsymbol{U}_2^{(d)} \otimes \boldsymbol{U}_1^{(d)})^\top \boldsymbol{y}. \tag{18}$$

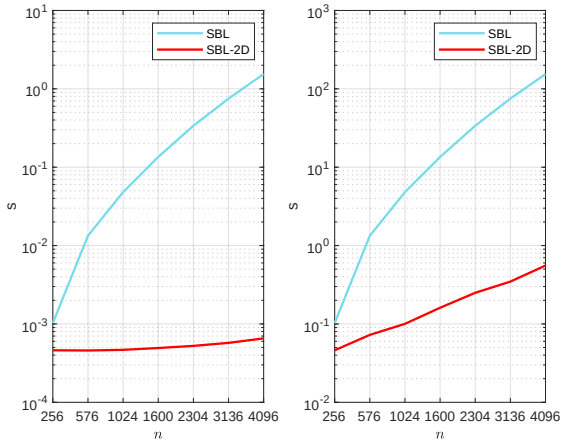

Figure 1: Timpii de execuție (secunde) pentru SBL și SBL-2D. Stânga: timp per iterație. Dreapta: timp total.

Pasul E este finalizat prin transformarea expresiei matricii de covarianță (7)

$$\boldsymbol{\Sigma}_x^{(d)} = \boldsymbol{B}_2^{(d)} \otimes \boldsymbol{B}_1^{(d)} - (\boldsymbol{E}_2^{(d)} \otimes \boldsymbol{E}_1^{(d)}) \cdot \operatorname{diag}(\boldsymbol{t}^{(d)}) \cdot (\boldsymbol{E}_2^{(d)} \otimes \boldsymbol{E}_1^{(d)})^\top. \tag{19}$$

Până la acest punct am prezentat elementele necesare pasului E al algoritmului EM. Pasul M presupune actualizarea valorilor hiperparametrilor și estimarea zgomotului. Adaptarea acestui pas la problema separabilă devine

$$\beta_i^{(d)} \leftarrow \frac{1}{n_{3-d}} \sum_{\substack{\ell=1 \\ c_d(\ell)=i}}^{n} \left((\Sigma_x)_{\ell,\ell}^{(d)} + (\mu_\ell^{(d)})^2 \right). \tag{20}$$

În final, zgomotul este estimat separat pentru cele două procese

$$(\sigma^{(d)})_{(k+1)}^2 = \frac{1}{m_1 m_2} \left(\|\boldsymbol{y} - (\boldsymbol{D}_2 \otimes \boldsymbol{D}_1)\boldsymbol{\mu}^{(d)}\|^2 + (\sigma^{(d)})_{(k)}^2 \sum_{\ell=1}^{n} \left(1 - \frac{(\Sigma_x)_{\ell,\ell}^{(d)}}{\beta_{c_d(\ell)}^{(d)}} \right) \right)^{1/2}. \tag{21}$$

Prima etapă a algoritmului SBL-2D este astfel finalizată prin identificarea suportului reprezentării. Cea de-a doua etapă presupune aplicarea algoritmului standard SBL pe suportul restrâns, cu scopul de a calcula elemenetele reprezentării. Această abordare în doi pași reduce costul computațional, deoarece problema pe care se aplică SBL are o dimensiune considerabil mai mică decât cea inițială. Figura 1 prezintă timpii de execuție pentru amele soluții. Testele sunt efectuate pe date sintetice având nivelul de sparsitate reală $s = 5$ și afectate de zgomot sub raportul semnal-zgomot SNR $= 40$.

Rezultatele arată un nivel de acuratețe comparabil cu cel obținut de metoda SBL și în multe cazuri semnificativ mai bun decât cele obținute cu alte metode

de calcul al reprezentării, adaptate pentru cazul 2D. Testele de performanță sunt realizate pe un set de semnale sintetice și pe un set de imagini reale. În cazul datelor sintetice, testele includ diferite nivele de sparsitate și zgomot și evaluează eroarea de aproximare a semnalelor (RMSE) și eroarea de reprezentare (comparând reprezentarea reală cu cea estimată). De asemenea, sunt realizate teste de convergență ce arată reducerea suportului reprezentării cu fiecare iterație, evoluția celor două măsuri de eroare, cât și numărul de fals pozitive și fals negative în suportul identificat.

3 Adaptarea Dimensiunii Dicționarelor pentru Dicționare Separabile

Atunci când nivelul de sparsitate nu este cunoscut sau estimat prin una din strategiile descrise anterior, este de regulă stabilit în funcție de dimensiunea dicționarului. Însă dacă în cazul sparsității este posibilă o decizie informată asupra valorii (pentru anumite semnale), în ceea ce privește dimensiunea dicționarului, astfel de informații sunt rareori disponibile. În practică, valoarea dimensiunii dicționarului este aleasă în funcție de dimensiunea semnalului. Alternativ, se poate opta pentru strategii adaptive în vederea determinării dimensiunii optime.

Oricare ar fi criteriul de optimalitate utilizat, aceste strategii presupun fie inițializarea unui dicționar de dimensiuni reduse, fie relativ mari și adăugarea și/sau eliminarea atomilor pe măsură ce procesul de antrenare avansează, în vederea obținerii unui model (dicționar) potrivit. Mai precis, optimalitatea dimensiunii este evaluată o dată la câteva iterații și dimensiunea dicționarului e modificată în mod corespunzător. Cu toate acestea, adaptarea dimensiunii poate interfera cu procesul de învățare, datorită faptului că atomii nou adăugați necesită un număr de iterații pentru a converge către direcții reprezentative pentru semnal.

O soluție pentru evaluarea modelului o reprezintă principiul Minimum Description Length (MDL) pentru selecția modelelor, ce se bazează pe presupunerea că semnalele conțin corelații ce permit o reprezentare parcimonioasă. Prin urmare, un model este optim în măsura în care poate comprima datele fără a reduce performanțele în aproximarea semnalelor. Abordarea include criterii de teoria informației (ITC) ce evaluează compromisul între acuratețea modelului și complexitatea acestuia. În particular, criteriul Renormalized Maximum Likelihood (RNML) permite o formulare (denumită Extended-RNML) adecvată problemei de estimare a dimensiuni optime a dicționarului [6].

Contribuția noastră implică adaptarea ERNML și a schemei de modificare a dimensiunii la modelul separabil [7]. Datorită echivalenței dintre problema originală DL (1) și cea separabilă (3), criteriul ERNML nu necesită reformulări fundamentale. Adaptările necesare se referă la ajustări ale dimensiuni și, având în vedere că existența a două dicționare, ajustarea numărului de parametri. Mai exact, dimensiunea semnalului este $T_{2D} = m_1 m_2 N$ iar numărul de parametri $\mathrm{NoP}_{2D} = sN + (m_1 - 1)n_1 + (m_2 - 1)n_2$, deoarece include gradele de libertate

ale fiecărui dicționar. Astfel, în cazul separabil, criteriul devine

$$\begin{aligned}\mathrm{ERNML}_{2D} = (T_{2D} - \mathrm{NoP}_{2D}) \log \frac{\mathrm{RMSE}^2}{T_{2D} - \mathrm{NoP}_{2D}} \\ + \mathrm{NoP}_{2D} \log \frac{\|\boldsymbol{D}_1 \boldsymbol{X} \boldsymbol{D}_2^\top\|_F^2}{T_{2D} \cdot \mathrm{NoP}_{2D}} \\ + \log\left[\mathrm{NoP}_{2D}(T_{2D} - \mathrm{NoP}_{2D})\right] + 2N \log \binom{n_1 n_2}{s}.\end{aligned} \tag{22}$$

Adaptarea dimensiunii dicționarului presupune în [6] evaluarea iterativă a compromisului complexitate-performanță pentru un număr n_{cand} de modele-candidați, anume a unor dicționare de dimensiuni diferite. Urmând principiul lui Occam, strategia înclină spre evaluarea dicționarelor de dimensiuni mici, ceea ce prezintă de asemenea avantajul unui cost computațional redus al criteriului. Este luat, însă, în calcul și cazul în care dimensiunea optimă este mai mare. Pe scurt, soluția testează dacă $\boldsymbol{D}$ este supradimensionat cu atomi nefolositori. Prin urmare, la fiecare $iter_{adapt}$ iterații, atomii dicționarului curent sunt ordonați în funcție de puterea reprezentării. Puterea unui atom j este definită în [1]

$$P_j = \|\boldsymbol{x}_j^\top\|_2^2. \tag{23}$$

Modelele-candidați sunt formate prin excluderea atomilor având puterea cea mai mică, astfel că modelul cel mai parcimonious va avea $n_{current} - n_{cand} + 1$ atomi. Reprezentarea $\boldsymbol{X}$ este apoi calculată pentru fiecare model pentru a evalua criteriul ERNML. Dicționarul optim este acela pentru care valoarea criteriului este minimă. Cazul în care minimul se obține pentru dicționarul curent este interpretat ca o indicație a faptului că modelul este subdimensionat, deci este necesară adăugarea unor atomi noi.

În cazul dicționarelor separabile, noțiunea de putere a reprezentării se referă la combinația de atomi $\boldsymbol{D}_{1,i}$ and $\boldsymbol{D}_{2,j}$. Mai exact, mărimea are acum forma

$$P_{2D} = \sum_{k=1}^{N} X_{i,j,k}^2. \tag{24}$$

Prin urmare, este necesară ordonarea combinațiilor de atomi după criteriul erorii, rezultând câte o listă de atomi pentru D_1 și D_2. Vor fi acum n_{cand}^2 candidați, deoarece n_{cand} atomi pot fi excluși din fiecare dicționar. Valoarea criteriului este apoi calculată pentru fiecare combinație de dimensiuni de dicționare. Ajustarea dimensiunii se realizează separat pentru fiecare dicționar, operând adăugările sau eliminările de atomi indicate de valoarea minimă a ERNML. Cele două dicționare $\boldsymbol{D}_d$ pot avea dimensiuni diferite, ceea ce se traduce printr-o flexibilitate a modelului, dat fiind faptul că tiparele prezente într-o dimensiune a semnalului (spre exemplu liniile din $\boldsymbol{y}$) pot fi diferite în ce privește dimensiunea și complexitatea de cele din cea de-a doua dimensiune (coloanele $\boldsymbol{y}$), necesitând astfel un număr diferit de atomi în D_1 și D_2.

Propunem de asemenea o soluție care minimizează interferența dintre modificarea dimensiunii și procesul de învățare, care implică utilizarea lui ERNML_{2D} doar ca indicator al schimbării. Dimensiunea optimă este calculată netezind valorile obținute de criteriu în ultimele ws iterații, pentru a evita modificările abrupte de dimensiune, ce pot determina prezența unor atomi antrenați insuficient. Soluția utilizează un filtru medie alunecătoare cu fereastră de dimensiune ws. Schema de adaptare a dimensiunii este prezentată în Algoritmul 1. Rezultatele arată o performanță bună în recuperarea dimensiunilor dicționarelor, pentru diferite configurații de dimensiuni, nivele diferite de sparsitate și zgomot. În mod special, dimensiunea este rar subestimată, cu atât mai puțin dacă cea reală este mică.

4 Anomaly Detection

Problema detecției de anomalii (AD) poate fi interpretată ca o clasificare binară, în care una din clase reprezintă semnalele normale, iar cealaltă anomaliile. Contribuțiile noastre utilizează DL pentru a rezolva această problemă pentru trei tipuri de aplicații: identificarea de malware, detecția de fraude financiare și problema mai generală a detecției de anomalii pe grafuri.

Detecția de malware este o problemă de dimensiuni mari, având în vedere multitudinea de aplicații software existente. Costurile computaționale pe care le implică o astfel de problemă pot fi reduse dacă se utilizează algoritmi online, ce permit antrenarea modelului succesiv pe sub-seturi ale datelor originale, fără a compromite acuratețea. Un alt aspect relevant îl constituie dinamica domeniului, în care tipuri noi de malware sunt create în mod constant, pentru a se sustrage soluțiilor anti-virus. Este necesar ca modelele să fie deschise față de această adaptabilitate, ceea ce se traduce prin abilitatea de a identifica tipuri de malware pentru care nu există exemple prealabile. Prin urmare, modelele nesupervizate sunt preferabile.

Observațiile de mai sus sunt valabile și în cazul detecției de fraude financiare, poate chiar într-o măsură mai mare. Acestea aduc însă o problemă în plus. Tranzacțiile sunt modelate cel mai firesc ca legături între două noduri - entitățile financiare. Prin urmare, un model potrivit de a reprezenta astfel de date este graful, datorită posibilității de a reprezenta interdependențele tranzacțiilor. De obicei, atât nodurile cât și muchiile gragului sunt adnotate cu informații ca identitatea, suma tranzacționată sau valuta. Așadar, datele conțin atât informații numerice, cât și relaționale. Graful tranzacțiilor este în special relevant în cazul schemelor de spălare de bani, în care tranzacțiile individuale pot părea legitime, ele apărând ca fraude doar dacă se consideră contextul mai larg al grafului în care sunt legate.

Motivată de problema identificării schemelor de spălare de bani și a altor fraude financiare, lucrarea noastră [8] prezintă un studiu al metodelor de detecție de anomalii destinate semnalelor de tip graf.

Cadrul general DL poate fi extins la problema de clasificare, mai ales în scenarii de antrenare supervizată. Scopul este acela de a antrena un dicționar în

Algorithm 1: ITC-ADL-2D

Data:
semnalele, $\boldsymbol{Y} \in \mathbb{R}^{m \times N}$
sparsitatea, s
numărul de iterații DL, K
pasul de iterații pentru adaptarea dimensiunii, K_{adapt}
numărul de modele-candidat, n_{cand}
dimensiunea minimă a unui dicționar n_{min}
fereastra mediei alunecătoare, ws

Result: dicționarele cu dimensiune optimă, $\boldsymbol{D}_1$ și $\boldsymbol{D}_2$

Inițializează $\boldsymbol{D}_1$ și $\boldsymbol{D}_2$ cu dimensiunile n_d
for $k = 1 : K_{adapt} : K$ **do**
    Antrenează K_{adapt} iterații DL-2D pentru a obține dicționarele actualizate $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ și reprezentarea $\boldsymbol{X}$ având sparsitate s
    Ordonează separat atomii din $\boldsymbol{D}_1$ și $\boldsymbol{D}_2$ în funcție de puterea reprezentării
    Calculează criteriul ERNML_{2D} pentru cei n_{cand}^2 candidați utilizând (22) și returnează dimensiunile $n_{d,ITC}$ corespunzătoare valorii minime a criteriului ERNML_{2D}
    Calculează $n_{d,opt}$ filtrând valorile $n_{d,ITC}$ pe o fereastră de dimensiune ws
    Aplică $\boldsymbol{D}_d = \texttt{AdjustSize}(\boldsymbol{D}_d, n_{d,opt}, n_d)$, unde $d \in \{1, 2\}$ pentru ambele dicționare $\boldsymbol{D}_1$ și $\boldsymbol{D}_2$
Antrenează K_{adapt} iterații DL-2D cu dimensiunile optime $n_{d,opt}$
Function $\boldsymbol{D}_d = \texttt{AdjustSize}(\boldsymbol{D}_d, n_{d,opt}, n_d)$
    if $n_{d,opt} - n_d > 0$ **then**
        Dimensiunea curentă $n_d = n_{d,opt}$
        Adaugă $n_{d,opt} - n_d$ atomi la $\boldsymbol{D}_d$
    else if $n_{d,opt} - n_d < 0$ **then**
        Dimensiunea curentă $n_d = \min(n_{d,opt}, n_{min})$
        Elimină atomi din $\boldsymbol{D}_d$ astfel încât dimensiunea acestuia să fie n_d
    else
        Dimensiunea curentă $n_d = n_{d,opt} + 1$
        Adaugă 1 atom la dicționar $\boldsymbol{D}_d$

care atomii sunt specializați pentru a descrie semnalele din fiecare clasă. În plus, este de dorit ca reprezentările semnalelor dintr-o clasă să difere semnificativ de cele din alte clase. Aceste două obiective adiționale pot fi impuse sub forma unor termeni de regularizare în problema DL, astfel încât $\boldsymbol{X}$ să reprezinte etichetele claselor și să respecte alocarea atomilor. Așadar, noua problemă de optimizare conține obiectivul obișnuit de reconstrucție a semnalelor, discriminarea între

clase și consistența atomilor (alocarea atomi-clase)

$$\min_{\boldsymbol{D},\boldsymbol{W},\boldsymbol{A},\boldsymbol{X}} \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2+\alpha\|\boldsymbol{H}-\boldsymbol{W}\boldsymbol{X}\|_F^2+\beta\|\boldsymbol{Q}-\boldsymbol{A}\boldsymbol{X}\|_F^2 \tag{25}$$

unde $\boldsymbol{H}\in\mathbb{R}^{c\times N}$ reprezintă matricea etichetelor, $\boldsymbol{W}\in\mathbb{R}^{c\times n}$ clasificatorul, $\boldsymbol{Q}\in\mathbb{R}^{n\times N}$ matricea de alocare a atomilor, $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ este dicționarul ce impune consistența atomilor asupra reprezentării, iar c este numărul de clase. Soluția se numește Label Consistent K-SVD (LC-KSVD) [9], datorită faptului că poate fi rescrisă în termenii problemei clasice DL și poate fi prin urmare rezolvată cu algoritmi standard, cum este K-SVD. În particular, (25) este echivalentă cu

$$\min_{\boldsymbol{D},\boldsymbol{W},\boldsymbol{A},\boldsymbol{X}} \left\| \begin{bmatrix} \boldsymbol{Y} \\ \sqrt{\alpha}\boldsymbol{H} \\ \sqrt{\beta}\boldsymbol{Q} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D} \\ \sqrt{\alpha}\boldsymbol{W} \\ \sqrt{\beta}\boldsymbol{A} \end{bmatrix} \boldsymbol{X} \right\|_F^2 . \tag{26}$$

Abordările online ale problemei de detecție de anomalii utilizând DL utilizează, în mod obișnuit, o serie de parametrii euristici pentru a controla încrederea în rezultatul clasificării. Cu toate că această soluție aduce beneficii de acuratețe, nu este potrivită aplicațiilor avute în vedere. Atât fraudele financiare, cât și aplicațiile malware presupun un grad înalt de noutate: este de așteptat ca noi scheme de fraudă să fie create pe măsură ce cele vechi sunt detectate de eforturile instituționale împotriva spălării banilor și, în egală măsură, noi tipuri de malware sunt dezvoltate pe măsură ce programele antivirus le identifică pe cele existente.

Propunem o metodă semi-supervizată (numită Tolerant Online Discriminative DL with Regularization - TODDLeR) pentru a rezolva acest compromis. Aceasta presupune o etapă off-line de pre-antrenare, unde dicționarul este antrenat pe un set etichetat de semnale, de dimensiuni reduse. Această etapă poate fi rezolvată cu metode standard de clasificare DL, cum este LC-KSVD descris mai sus. Dicționarul este apoi utilizat pentru a inițializa faza nesupervizată, online, unde antrenarea lui $\boldsymbol{D}$ se realizează cu fiecare nou semnal. Soluția și rezultatele descrise mai jos reprezintă contribuția noastră, publicată în [10]. Metoda este bazată pe un algoritm existent, [11], însă spre deosebire de acesta, utilizează toate semnelele pentru a antrena modelul. Etichetarea eronată poate influența negativ procesul de antrenare, de aceea ne dorim ca semnalele să nu poată modifica în mod drastic modelul curent, în lipsa posibilității de a evalua aceste modificări. Prin urmare, soluția presupune adăugarea unor termeni de regularizare pentru clasificator și matricea de consistență a etichetelor, pentru a controla schimbările asupra acestora. Problema (26) devine astfel

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{W},\boldsymbol{A}} & \|\boldsymbol{y}-\boldsymbol{D}\boldsymbol{x}\|_2^2+\alpha\|\boldsymbol{h}-\boldsymbol{W}\boldsymbol{x}\|_2^2+\beta\|\boldsymbol{q}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \\ & +\lambda_1\|\boldsymbol{W}-\boldsymbol{W}_0\|_F^2+\lambda_2\|\boldsymbol{A}-\boldsymbol{A}_0\|_F^2 . \end{aligned} \tag{27}$$

În formularea de mai sus, am adoptat abordarea online: în locul întregii matrici de semnale $\boldsymbol{Y}$, soluția este căutată pentru fiecare semnal individual $\boldsymbol{y}$, având vectorul de etichete corespunzător $\boldsymbol{h}$ și un vector asociat pentru alocarea

atomilor $\boldsymbol{q}$. Prin $\boldsymbol{W}_0$ and $\boldsymbol{A}_0$ notăm valorile curente ale dicționarelor, antrenate pe semnalele anterioare. Valorile actualizate ale $\boldsymbol{W}$ și $\boldsymbol{A}$ pot fi obținute fixând restul variabilelor în (25) și rezolvând obiectivele corespunzătoare

$$f(\boldsymbol{W}) = \|\boldsymbol{h} - \boldsymbol{W}\boldsymbol{x}\|_2^2 + \lambda_1 \|\boldsymbol{W} - \boldsymbol{W}_0\|_F^2 , \tag{28}$$

$$g(\boldsymbol{A}) = \|\boldsymbol{q} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda_2 \|\boldsymbol{A} - \boldsymbol{A}_0\|_F^2 . \tag{29}$$

Egalând cu 0 gradienții funcțiilor de mai sus obținem următoarele soluții

$$\boldsymbol{W} = (\boldsymbol{h}\boldsymbol{x}^T + \lambda_1 \boldsymbol{W}_0)(\boldsymbol{x}\boldsymbol{x}^T + \lambda_1 \boldsymbol{I})^{-1}, \tag{30}$$

$$\boldsymbol{A} = (\boldsymbol{q}\boldsymbol{x}^T + \lambda_2 \boldsymbol{A}_0)(\boldsymbol{x}\boldsymbol{x}^T + \lambda_2 \boldsymbol{I})^{-1}. \tag{31}$$

Metoda este testată pe două baze de date de malware și o bază de date de fraude financiare. Rezultatele arată o acuratețe de clasificare comparabilă și în unele cazuri îmbunătățită față de alte metode DL online de clasificare.

Cea de-a doua contribuție la problema AD reprezintă o metodă nesupervizată în care semnalele normale sunt filtrate în mod progresiv [12]. Ipoteza ce stă la baza metodei o reprezintă faptul că în astfel de aplicații numărul semnalelor normale îl depășește cu câteva ordine de mărime pe cel al anomaliilor, astfel încât acestea sunt mai bine reprezentate de model. Prin urmare, eroarea de reconstrucție poate reprezenta un criteriu de diferențiere între cele două clase. Soluția ia în considerare și faptul că, pe măsură ce semnalele normale sunt filtrate, acest dezechilibru se atenuează. Propunem o structură de dicționar compozit, în care noile modele, obținute la fiecare iterație, sunt combinate cu cele existente deja, astfel încât se evită supra-antrenarea lui $\boldsymbol{D}$ pe anomalii.

O alternativă la criteriul de eroare o reprezintă proprietățile fiecărui atom, pe care le utilizăm pentru a obține informații despre clasa fiecărui semnal. Un atom poate fi caracterizat în funcție de numărul de semnale în a căror reprezentare se regăsește. Denumim această proprietate utilitatea atomului j, $U_j = \|\boldsymbol{x}_j^\top\|_0$. Așadar, o altă opțiune de a filtra progresiv semnalele, presupune restrângerea setului de posibile anomalii $\mathcal{A}$ la semnalele reprezentate de atomii $\boldsymbol{d}_j$ având utilitatea $U_j < N_a$,

$$\mathcal{A} = \{\, \boldsymbol{y}_k \mid x_{j,k} \neq 0 \wedge U_j < N_a, \forall j, k \,\} . \tag{32}$$

Cu alte cuvinte, este de așteptat ca atomii utilizați de mai puțin de N_a semnale, să fie anomalii, unde N_a reprezintă numărul de anomalii din baza de date.

Utilizând ambele metode se obțin rezultate bune în ce privește numărul de fals pozitive, în timp ce numărul de fals negative este de asemenea menținut la o valoare mică, ceea ce sugerează că pot fi utilizate pentru a reduce dezechilibrul bazelor de date.

O altă contribuție folosește de asemenea proprietatea de utilitate a atomilor, împreuna cu aceea de putere, într-o soluție în care dicționarul este structurat pe mai multe scale, cu scopul de a investiga măsura în care aceste caracteristici ale atomilor pot constitui indicatori ai clasei semnalelor. Rezultatele intermediare

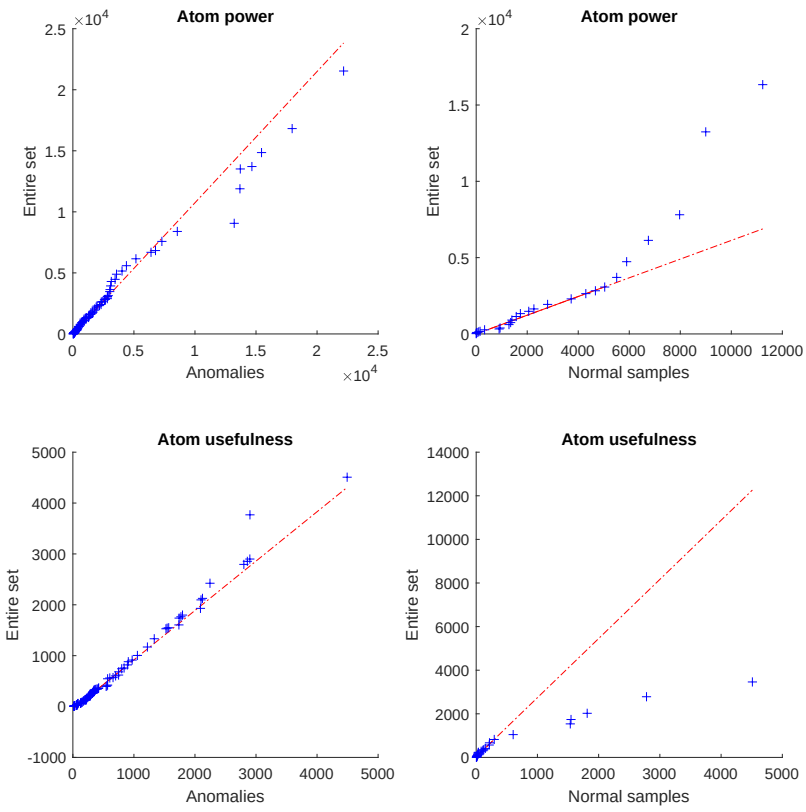

Figure 2: Graficul Quantile-Quantile - distribuțiile puterii și utilității atomilor pentru clasa semnalelor normale și anomaliilor

arată că problema standard DL, chiar fără a adăuga constrângerile legate de clasificare, poate oferi informații despre clasele semnalelor dacă se utilizează aceste proprietăți ale atomilor pentru a descrie semnalele. Experimentele arată că distribuțiile celor două mărimi diferă semnificativ între datele normale și anomalii. Figura 2 prezintă graficul Quantile-Quantile ce ilustrează această comparație.

Ultima contribuție, denumită Separable Laplacian Classification [13] exploatează structura grafului pentru a identifica topologii anormale. Aceasta utilizează modelul separabil, iar semnalele sunt reprezentate în funcție de matricea Laplacian. Strategia presupune, așadar, utilizarea acestor matrici ce descriu structura grafului pentru a învăța tipare de conectivitate specifice fiecărei clase de grafuri. Soluția este testată pe un set de date sintetice reprezentât semnale de tip graf, în care anomalii diferă structural de restul datelor. Tiparele

testate sunt de tip inel și clică, datorită prevalenței acestora în aplicațiile de detecție de fraudă. Semnalele de antrenare provenind din fiecare clasă sunt utilizate separat pentru a antrena câte o pereche de dicționare. Clasificarea unui nou semnal, de test, presupune determinarea perechii pentru care eroarea de reprezentare a semnalelui este minimă. Este investigat de asemenea efectul dimensiunii dicționarelor asupra performanțelor metodei în identificarea de tipare circulare.

References

[1] B. Dumitrescu, P. Irofti, Dictionary Learning Algorithms and Applications, Springer International Publishing, 2018 (2018).

[2] M. Tipping, Sparse Bayesian Learning and the Relevance Vector Machine, Journal of Machine Learning Research 1 (2001) 211–244 (2001).

[3] D. P. Wipf, B. D. Rao, Sparse Bayesian learning for basis selection, IEEE Transactions on Signal Processing 52 (8) (2004) 2153–2164 (2004).

[4] A. Băltoiu, B. Dumitrescu, Sparse bayesian learning algorithm for separable dictionaries, Digital Signal Processing 111 (2021) 102990 (2021).

[5] O. Stegle, C. Lippert, J. Mooij, N. Lawrence, K. Borgwardt, Efficient inference in matrix-variate Gaussian models with iid observation noise, in: Proceedings of the 24th Neural Information Processing Systems Conference, 2011, pp. 630 – 638 (2011).

[6] B. Dumitrescu, C. D. Giurcăneanu, Adaptive-Size Dictionary Learning Using Information Theoretic Criteria, unpublished document (2019).

[7] A. Baltoiu, B. Dumitrescu, Size adaptation of separable dictionary learning with information-theoretic criteria, in: 2019 22nd International Conference on Control Systems and Computer Science (CSCS), 2019, pp. 7–11 (2019).

[8] P. Irofti, A. Băltoiu, A. Pătrașcu, Fraud detection in networks, in: Enabling AI applications in Data Science, Springer, 2020, pp. 517–536 (2020).

[9] Z. Jiang, Z. Lin, L. Davis, Learning A Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD, in: IEEE Conf. Computer Vision and Pattern Recognition, 2011, pp. 1697–1704 (2011).

[10] P. Irofti, A. Băltoiu, Malware identification with dictionary learning, in: 27th European Signal Processing Conference, 2019, pp. 1–5 (2019).

[11] S. Matiz, K. Barner, Label consistent recursive least squares dictionary learning for image classification, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 1888–1892 (2016).

[12] P. Irofti, A. Băltoiu, Unsupervised dictionary learning for anomaly detection, in: International Traveling Workshop on Interactions Between Sparse Models and Technology, 2020, pp. 1–3 (2020). arXiv:2003.00293.

[13] A. Băltoiu, A. Pătrașcu, P. Irofti, Graph anomaly detection using dictionary learning, in: The 21st World Congress of the International Federation of Automatic Control, 2020, pp. 3551–3558 (2020).