Nr. 798 /28.06.2018

RAPORT DE CERCETARE INTERMEDIAR NR. 1
PROIECT „SMART BUILDING. SMART CITY."
ACADEMIA OAMENILOR DE ȘTIINȚĂ DIN ROMÂNIA (AOSR)

Modelarea prin algoritmi de învățare statistică a subsistemului de ventilație dintr-o clădire inteligentă

Coordonator proiect: Dr.-Ing. Vladimir Tanasiev

Autor: Dr.-Ing. Grigore Stamatescu

Iunie 2018

Rezumat

Considerând dezvoltările recente din domeniul monitorizării și conducerii clădirilor prin rețele de dispozitive interconectate inteligente, gestionarea fluxurilor bogate de date asociate a devenit o provocare importantă din punct de vedere științific și aplicativ. În multe situații, aplicabilitatea metodelor clasice de identificare de sistem sau a modelelor de aproximare de tipul grey-box pot să nu fie fezabile sau adecvate. În acest raport este discutată și ilustrată pe seturi de date reale aplicabilitatea modelării black-box intrare-ieșire, prin tehnici de data mining, a subsistemelor aferente funcției de ventilație a sistemului HVAC dintr-o clădire. Studiul de caz prezentat include modelarea a patru centrale de tratare a aerului instalate într-o clădire modernă din campusul Universității Politehnica din București.

Cadrul de prelucrare de date și învățare conține doi pași: fluxurile brute de date sunt comprimate prin metoda de aproximare prin agregare simbolică (SAX), segmentele agregate rezultante fiind folosite în a doua etapă pentru algoritmi de clasificare cu mașini de vectori suport (SVM). Rezultatele se pot dovedi utile pentru discriminarea modului de operare a unităților CTA individuale în regimuri de funcționare diverse și pot fi folosite în aplicații de nivel înalt pentru detecția defectelor sau creștere eficienței energetice.

În acest context, ne adresăm aplicării acestor tehnici avansate de prelucrări de date în domeniul sistemelor de automatizare ale clădirilor (BAS). Într-o structură BAS modulară, sistemul de încălzire, răcire și ventilație (HVAC) joacă un rol dominant în consumul energetic global cât și pentru condițiile de confort percepute de utilizatorii finali, în special prin evaluarea subiectivă a confortului termic interior. În funcție de amplasarea geografică și alte constrângeri locale și tehnice, structura sistemelor HVAC poate varia. Rolul centralelor de tratarea a aerului în sistemul HVAC este de a asigura funcția de ventilație prin aportul cantității optime de aer proaspăt, filtrarea acestuia și condiționarea acestuia prin încălzire sau răcire, pe un domeniu restrâns. Prin recircularea aerului extras din clădire, unitățile CTA contribuie și la eficiența energetică a acestora.

Principalul obiectiv al acestui raport în contextul proiectului de cercetare „Smart Building. Smart City." este ilustrarea aplicării unei metodologii de data mining pe datele colectate de la sistemul de ventilație al unei clădiri inteligente. Considerăm abordarea aleasă ca fiind reprezentativă și adaptabilă și pentru alte subsisteme ale unei clădiri. Sunt vizate în special clădirile comerciale de dimensiuni mari unde impactul absolut al unor astfel de tehnici poate fi semnificativ.

2. Stadiul actual al cercetărilor

Metodele din știința datelor ce pot fi aplicate pentru modelarea și conducerea proceselor din clădiri inteligente sunt identificate și descrise pe larg de către [1]. Scopul principal în acest caz este sprijinul profesioniștilor specializați în managementul energiei prin instrumente de suport al deciziei performante. Procesul de data mining este prezentat în conexiune cu natura specifică a aplicației și este organizat începând de la datele brute colectate, preprocesare, prelucrare, modelare, validare model și predicție. Algoritmii și instrumentele adecvate pentru astfel de sarcini sunt și ele discutate, atât produse open-source cât și comerciale. În [2] este propusă o arhitectură de predicție pe două niveluri pentru predicția consumul energetic pe termen foarte scurt, orizont de predicție de câteva ore până la o zi, și pe termen scurt, o zi până la câteva săptămâni. Prelucrarea datelor utilizează o arhitectură lambda ce împarte setul disponibil în două niveluri. Fiecare nivel aplică un algoritm diferit bazat pe modelarea ARIMA, în funcție de fereastra de timp necesară cu prelucrare în timp real pentru predicția la nivel de oră sau prelucrare mai lentă pentru predicția la nivel de zi.

Seturile de date de referință la nivel de clădiri [6] sunt foarte importante pentru evaluarea comparativă a metodelor și algoritmilor propuși și obținerea unor rezultate reproductibile. Autorii lucrării descriu o bază de date de dimensiuni mari ce conține consumurile de energie pe durata unui an pentru 507 clădiri comerciale, în special cu destinație academică. Seriile de timp sunt disponibile gratuit și reprezintă o referință foarte utilă pentru aplicarea și compararea unor metode de data mining specifice pentru modelarea eficienței energetică și extragerea trăsăturilor caracteristice. Provocările tehnice pentru obținerea de date de bună calitate în timp util pentru modelare și ținând cont de existență unor soluții BAS eterogene de la o serie de furnizori diferiți și tehnologiile diverse de colectare, stocare și regăsire a datelor sunt discutate pe larg. Controlul predictiv pentru menținerea confortului termic în clădiri este aplicat în [7]. Indexul de confort optimal este atins printr-o funcție de cost ce depinde atât de confortul ocupanților cât și de costul energiei.

Analizele de tipul big data pentru consumul de energie electrică al unui oraș inteligent sunt evidențiate în [8]. Autorii utilizează algoritmi de inteligență computațională pentru modelarea consumului a opt clădiri universitare. Rezultatul tangibil este reprezentat de politici offline de optimizare a energiei utilizate în campus. În [9] este prezentată o aplicație ce folosește arbori de decizie pentru estimarea gradului de ocupare în clădirile de birouri. Modelarea și estimarea ocupării este un obiectiv critic în clădirile inteligente dat fiind faptul că predicția îmbunătățită a acesteia are impact direct asupra strategiei de condiționare HVAC a clădirii, reducând risipa de energie. Din perspectiva detecției și identificării defectelor la senzori [10] sunt prezentare rezultatele strategiei de diagnoză BRIDGE într-un scenariu experimental dedicat. Considerând defectele senzorilor ca devieri de date, FDD poate detecta cu precizie bună condițiile anormale. [11] prezintă în detaliu procesul explicit de modelare a datelor pentru evaluarea clădirilor inteligente. Este descris un studiu de caz pentru predicția consumului energetic folosind tehnici precum rețelele neuronale cu regularizare bayesiană și random forests. Sunt considerate și tehnicile SVM dar în acest caz de testare rezultate obținute sunt mai slabe. În [12] SVM sunt aplicate cu succes pentru o problemă de regresie.

Figura 1 ilustrează sumar rolul metodelor de data mining la nivelul superior de decizie pentru scheme de control ierarhic bazat pe date. În studiul de caz prezentat fiecare unitate CTA include propriile bucle de reglare locale, în special temperatură și presiune relativă, ce respectă valorile de referință preluate de la sistemul BMS, în contextul unui program de

SAX reprezintă o extindere a metodei de aproximate agregată pe porțiuni (PAA) [15] prin atribuirea de simboluri segmentelor individuale identificate de PAA. Segmentele astfel determinate pot fi incorporate într-un model Markov pentru calculul probabilității de apariției a respectivelor secvențe pentru succesiuni de observații ulterioare. Conform descrierii metodei PAA, pornind de la o serie de timp X de lungime m, aceasta este aproximată cu un vector $\bar{X} = (\bar{x}_1, \ldots, \bar{x}_M)$ de lungime arbitrară $M \leq n$. Fiecare element din vectorul $\bar{x}_i$ este calculat ca:

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j$$

Asta înseamnă că reducem dimensionalitatea seriei de timp de la n la M eșantioane prin împărțirea inițială a datelor originale în M cadre de dimensiuni egale și ulterior calculăm valorile medii pentru fiecare cadru. Prin reasamblarea valorilor medii obținem o nouă secvență care este considerată transformarea (aproximarea) PAA a setului de date original. Referitor la considerațiile de calcul, complexitatea transformării PAA poate fi redusă de la O(NM) la O(Mm) cu m fiind numărul de cadre ca parametru al metodei. Metrica de distanță este definită ca:

$$D_{PAA}(\bar{X}, \bar{Y}) = \sqrt{\frac{n}{M}} \sqrt{\sum_{i=1}^{M} (\bar{x}_i - \bar{y}_i)}$$

S-a arătat că metoda PAA satisface condiția de mărginire inferioară și garantează lipsa rejecțiilor false, astfel încât:

$$D_{PAA}(\bar{X}, \bar{Y}) \leq D(X, Y)$$

1.2. Mașini cu vectori suport (SVM – Support Vector Machines)

Mașinile cu vectori suport [16] pentru probleme de clasificare extind problema clasificării cu discriminat liniar prin impunerea unei distanțe minime b>0 față de separatorul de clasă. În cazul unei probleme cu două clase avem astfel:

$$w \cdot x_k^T + b \geq +1 \quad \text{for} \quad y_k = 1$$

$$w \cdot x_k^T + b \leq -1 \quad \text{for} \quad y_k = 2$$

SVM caută soluția optimală prin minimizarea funcției obiectiv:

În cazul problemei de clasificare curente, reprezentată de identificarea corectă a datelor colectate online uneia dintre cele patru unități CTA, avem de-a face cu o problemă SVM multi-clasă, ca extindere a problemei binare descrise mai sus. Această este rezolvată printr-o abordre one-versus-one [17] ce presupune evaluarea tuturor perechilor de clasificatori dintre clasele țintă, cu k(k-1)/2 clasificatori. Rezultatul final este obținut prin numărarea voturilor individuale ale fiecărui clasificator pentru exemplul de test. Figura 2 ilustrează grafic abordarea propusă pentru etapele de prelucrare a datelor și învățare precum și perspectivele de utilizare în sisteme de suport al deciziei pentru operatori umani sau direct în bucle de reglare automată.

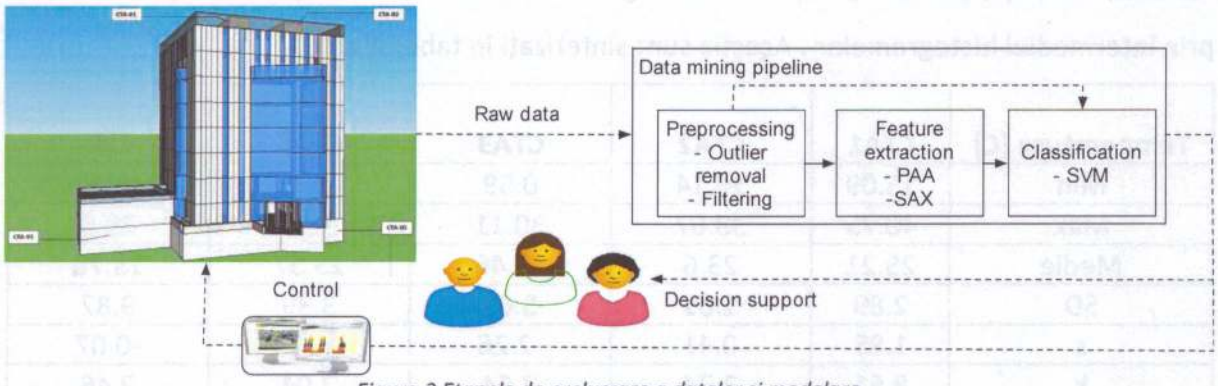

Figura 2 Etapele de prelucrare a datelor și modelare

4. Studiu de caz și rezultate

Datele colectate și utilizate în acest studiu provin de la cele patru unități CTA instalate în clădirea centrului de cercetări PRECIS al Universității Politehnica din București. Aceasta este o clădire de șapte etaje, circa 9000 mp suprafață utilă și utilizare mixtă prin laboratoare de cercetare, spații multifuncțională, săli de ședință precum și un auditoriu de 250 de locuri și birouri administrative, finalizată în anul 2016. Pentru monitorizarea și controlul diferitelor subsisteme este implementată o soluție software BMS furnizată de firma Honeywell și implementată pe un server central. Datele de interes sunt stocate într-o bază de date structurată de tip SQL și au fost colectate offline pentru studiu. De la fiecare unitate CTA avem acces la valorile de temperatură aferente aerului evacuat, celui introdus în clădire și cel recirculat, eșantionate la intervale de cinci minute. Suplimentar colectă valorile de referință pentru temperatura și umiditatea exterioară. Perioada de referință pentru studiu este întregul an 2017, mai exact perioada dintre 7 Ianuarie 2017 și 31 Decembrie 2017, un total de 359 de

În continuare sunt prezentate rezultatele obținute prin aplicarea celor doi pași din metodologia de data mining propusă. Seriile de timp preprocesate sunt reprezentate în formă agregată, inițial prin segmente numerice PAA urmate de segmente simbolice SAX. Unul dintre parametrii cheie ce determină performanța metodei este numărul de segmente pentru reprezentarea zilnică a fiecărei serii de timp și dimensiunea alfabetului în care aceste segmente sunt codificate de SAX. Un compromis între numărul de elemente de aproximare și granularitatea fiecărui element este prezentat. Configurația aleasă conduce la reducerea cu un factor de aproape 30 a cantității brute de date. Ulterior pe aceste date sunt antrenați și testați mai mulți clasificatori SVM. Pentru comparații relevante toate rezultatele sunt obținute pe un PC desktop cu procesor quad-core i7 3.6GHz și 16GB memorie RAM. Rezultatele sunt prezentate relativ la o referință comună sub forma unui clasificator svm cu funcție kernel gaussiană fină. O observație relevantă este că setul original de date conține informații redundante substanțiale având în vedere colectarea cu rată mare de eșantionare pentru procesele suport termice cu dinamică lentă, în timp ce reprezentarea agregată este posibil să piardă informații utile dar putem antrena mai rapid clasificatorii svm și testa mai multe configurații în același timp, putând îmbunătăți gradual un model care oferă inițial performanțe mai slabe.

Figura 4 prezintă rezultatele aplicării metodei SAX pe datele de temperatură de evacuare colectate pe tot parcursul anului 2017 de la CTA1 prin varierea celor doi parametri de proiectare: numărul de segmente *w* și dimensiunea alfabetului de simboluri *a*.

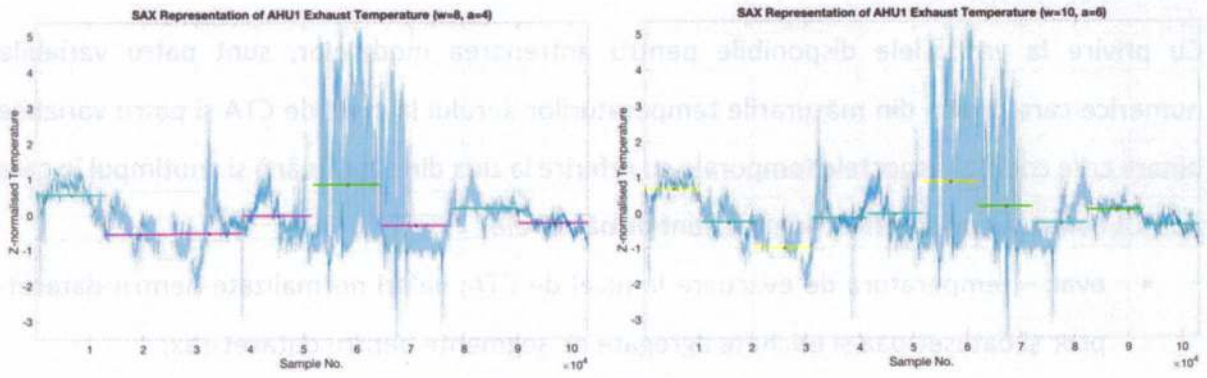

Figura 4 Rezultate SAX anuale cu diferiți parametri w și a

În Figura 5 sunt prezentate rezultatele SAX cu parametri aleși w=10 și a=6 pentru seriile de timp normalizate, la nivel de lună și de zi. Rezultate similare au fost obținute pentru toate cele

- ext – temperatura exterioară; aceeași pentru toate CTA;
- iswkn – variabilă binară care simbolizează dacă măsurarea a fost efectuată într-un weekend (1) sau nu (0);
- iswinter – variabilă binară care simbolizează dacă măsurarea a fost efectuată iarna (1) sau nu (0);
- issummer - care simbolizează dacă măsurarea a fost efectuată vara (1) sau nu (0);
- isshoulder - care simbolizează dacă măsurarea a fost efectuată într-un anotip intermediar, primăvara sau toamna, (1) sau nu (0);
- class – etichetele de clasă corespunzătoare celor patru unități de ventilație, numerotate cu 1, 2, 3 și 5.

Pe baza acestor date de intrare am antrenat modele de clasificare SVM multi-clasă și prezentăm rezultatele finale obținute. Din testele preliminare, particularitățile seturilor de date de intrare necesită separatoare de clasă mai complexe astfel încât sunt prezentate rezultate obținute prin utilizarea de funcții kernel cubice și gaussiene fine. Funcția kernel svm cubică are forma $k(x_j, x_k) = \left((x_j) \cdot (x_k)^T + 1\right)^3$. Funcția kernel svm gaussiană are forma $k(x_j, x_k) = exp\left(-\frac{\|x_j - x_k\|^2}{2\sigma^2}\right)^3$. În aplicația practică utilizăm un model gaussian fin ce folosește un factorul de scalare al funcției kernel $\sqrt{P}/4$ cu P numărul de variabile. Parametru de ajustare este factorul de scalare, determinat automat în cazul de față printr-o abordare heuristică aleatorie de subeșantionare. Sunt evaluați astfel în final șase clasificatori, câte doi pentru fiecare dintre cele trei seturi de antrenare. Rezultatele antrenării sunt prezentare în Tabelul 3 și ilustrate grafic în Figura 6.

SVM	Acuratețe [%]	AUC	Acuratețe [%]	AUC	Acuratețe [%]	AUC
kernel / set de date	dataset-proc		dataset-paa		dataset-sax	
Cubic	81.2	0.94	63.2	0.8675	70.9	0.91
Gauss	91.4	0.985	83.6	0.958	84.6	0.965

5. Concluzii

În acest raport a fost prezentat un studiu de caz întru argumentarea unei metodologii de data mining pentru modelarea subsistemului de ventilație al unei clădiri inteligente. Motivația principală derivă din necesitatea înțelegerii mai bune, cu date limitate, a dinamicii clădirii precum și a acțiunilor operatorului sistemului de automatizare și îmbunătățirea ulterioară a acestora. Cadrul prezentat are aplicabilitate generală și în afara studiul de caz ales pentru alte categorii de probleme legate de confortul utilizatorilor și/sau eficiența energetică în contextul mediului construit. Rezultatele prezentate sunt reaplicabile iar implementarea a fost realizată în mediul de programare tehnică MATLAB. Contribuția esențială a raportului este dată de modelele celor patru unități CTA din clădirea investigată ce pot fi folosite pentru detecția unor moduri de operare defectuoase sau ineficiente energetic.

Scopul principal a fost atins printr-o mai bună înțelegere a tiparelor de temperatură, ventilație și operaționale (acțiuni ale operatorului sistemului de automatizare al clădirii) și cercetările pot continua către ajustarea automată a referințelor de temperatură la nivel de CTA generate de regulile și modelele analitice învățate. Pentru operarea on-line, în vederea integrării rutinelor și modulelor software dezvoltate cu sistemul comercial de management al clădirii, se poate implementa o platformă middleware adecvată, cum este VOLTTRON [18] pe clădirea de interes. Cu referire la metodele de data mining și învățare se are în vedere alegerea unor funcții kernel îmbunătățire și identificarea optimă a hiperparametrilor de modelare. Există în prezent o serie de utilitare de nivel înalt dar și biblioteci eficiente ce permit îmbunătățirea semnificativă a performanțelor de modelare, inclusiv pe baza unor sisteme cloud.

[12] David M. Solomon, Rebecca Lynn Winter, A.G.B.R.N.A.L.L.W. Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression. Technical Report https://doi.org/10.7916/D85D90X7, Columbia University Academic Commons, 2011.

[13] Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery; ACM: New York, NY, USA, 2003; DMKD '03, pp. 2–11. doi:10.1145/882082.882086.

[14] Keogh, E.; Lin, J.; Fu, A. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. Proceedings of the Fifth IEEE International Conference on Data Mining; IEEE Computer Society: Washington, DC, USA, 2005; ICDM '05, pp. 226–233. doi:10.1109/ICDM.2005.79.

[15] Chakrabarti, K.; Keogh, E.; Mehrotra, S.; Pazzani, M. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. *ACM Trans. Database Syst.* **2002**, *27*, 188–228. doi:10.1145/568518.568520.

[16] Hastie, T.; Tishbirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer, 2009.

[17] Kressel, U.H.G. Advances in Kernel Methods; MIT Press: Cambridge, MA, USA, 1999; chapter Pairwise Classification and Support Vector Machines, pp. 255–268.

[18] Katipamula, S.; Haack, J.; Hernandez, G.; Akyol, B.; Hagerman, J. VOLTTRON: An Open-Source Software Platform of the Future. *IEEE Electrification Magazine* **2016**, *4*, 15–22. doi:10.1109/MELE.2016.2614178.