REZUMAT TEZA

A Wavelets Based Approach for Time Series Mining

Ing. Cristina Stolojescu

Aceasta teză are un obiectiv practic. Ea constă în găsirea unei soluții la următoarea întrebare: « Este posibilă identificare stațiilor de bază rău poziționate în topologia unei rețele WiMAX prin analiza traficului ?» Această întrebare este importantă pentru planificarea şi exploatarea unei rețele. Răspunsul ar putea fi o explicație a motivelor pentru care performanța rețelelor, măsurată în practică, este inferioară performanței estimate în faza de proiectare.

Abordarea selectată în această teză se bazează pe analiza unei baze de date concepută în urma monitorizării traficului dintr-o rețea WiMAX pe durata a opt saptămâni. Această rețea a fost dezvoltată de Alcatel Lucent Timişoara, România, şi este compusă din 67 de stații de bază (BS).

Având în vedere volumul mare de informații conținute în această bază de date, am preferat o abordare de data-mining. Ea se bazează pe metodologia CRISP-DM (CRoss Industry Standard Process for Data Mining), care este un model de proces data mining foarte des utilizat de experți pentru a rezolva diverse probleme ce apar în data mining. Metodologia CRISP-DM presupune urmatoarele etape:

1. Intelegerea problemei
2. Intelegerea datelor
3. Pregatirea datelor
4. Modelare
5. Evaluare
6. Implementarea solutiei finale

Prima etapa a unui proiect data mining este de a înțelege problema. Această etapă se implementează într-un mod iterativ, prin colaborare cu alte etape ale proiectului, ca de exemplu etapa de pregătire a datelor sau cu cea de evaluare. A doua etapă presupune înțelegerea datelor şi este în strânsă legătură cu prima etapă. O prima înțelegere a problemei implica o înțelegere preliminară a datelor. Cu aceste cunoştințe procesul de înțelegere a datelor este ameliorat şi, în consecința, procesul de înțelegere a problemei este de asemenea ameliorat.

Următoarea etapă este cea de pregătire a datelor. In general, datele brute sunt afectate de imperfecțiunea sistemelor de achiziție de date, de exemplu trasele conținute în baza de date furnizată de Alcatel Lucent conțin date lipsa. Acesta este motivul pentru care etapa de pregătire a

datelor este necesară într-un proiect de tip data mining. Această etapă presupune reprezentarea datelor într-o formă favorabilă pentru modelare. Etapa de modelare este una dintre cele mai importante etape din proiect pentru că presupune reprezentarea datelor în forma cea mai potrivită extragerii celor mai importanți parametrii pentru aplicația considerata. Etapa de evaluare permite aprecierea calității modelului ales. Ultima etapă a proiectului este cea de implementare a soluției finale.

Această metodologie este aplicată în lucrarea de față pentru extragerea de informații, interpretetarea lor şi propunerea de soluții.

Selecția traficului de comunicații, ca obiect de analiză în această teză, este justificată de următoarele motive:

1. traficul poate fi măsurat,
2. traficul poate şi trebuie să fie optimizat pentru a putea elabora strategii de creştere a performanței rețelei (în special în cazul comunicațiilor wireless, cum este cazul tehnologiei WiMAX).

Această selecție a orientat teza de față către cercetarea metodelor de analiză a seriilor de timp. Analiza seriilor temporale a devenit o provocare pentru mulți cercetători în ultimii ani. Originile sale pot fi găsite în cercetare în domeniul matematicii, dar astăzi analiza seriilor de timp este un domeniu multi-disciplinar, exploatând rezultatele obținute în matematică, procesarea statistică a semnalelor, data mining, inginerie etc. Acesta este motivul pentru care lucrarea de față are un caracter multi-disciplinar integrând competențele în informatică din departamentul LUSSI de la Telecom Bretagne, Brest, Franța, cu competențele în domeniul comunicațiilor din departamentul de comunicații al Facultății de Electronică şi Telecomunicații din cadrul Universitații "Politehnica" din Timişoara, România. Una dintre dificultățile majore ale analizei seriilor de timp cu lungimi mari (care corespund unui volum mare de date) este complexitatea mare de calcul. Această complexitate de calcul poate fi redusă prin reprezentarea datelor într-o formă favorabilă. Una dintre etapele metodologiei CRISP-DM, şi anume cea de pregătire a datelor, presupune reprezentarea datelor într-o formă mai favorabilă. O astfel de reprezentare poate fi obținută folosind unidişoarele. Transformata în undişoare discretă (DWT) se utilizează în analiza serilor temporale şi implică o complexitate redusă de calcul. Transformata în undişoare a fost utilizată pentru analiza seriilor de timp în multe lucrări în ultimii ani. Una dintre principalele proprietăți ale undişoarelor este că acestea sunt localizate în timp (sau spațiu), ceea ce le face potrivite pentru analiza semnalelor nestaționare (care conțin semnale tranzitorii şi structuri fractale).

Cadrul de cercetare asociat cu lucrarea de față are următoarele axe:

- Undişoare – o introducere este prezentată în Capitolul 1,
- Procesarea statistică a semnalelor – conceptele de bază sunt prezentate în Capitolul 2,
- Analiza seriilor temporale – realizată în Capitolul 3 şi Capitolul 4,
- Data-mining - realizată în Capitolul 3 (unde este evidențiată şi dezvoltată metodologia CRISP-DM) şi în Capitolul 4,
- Rețelele WiMAX – descrise în Capitolul 3 şi analizate în Capitolul 3 şi Capitolul 4.

Aşa cum s-a spus deja, scopul acestei teze este de a răspunde la întrebarea " Este posibilă identificare staţiilor de bază rău poziţionate în topologia unei reţele WiMAX prin analiza traficului? ". Presupunând că traficul asociat cu o staţie de bază prost poziţionată este mai greoi decât traficul asociat cu o staţie de bază bine poziţionată au fost elaborate două abordări pentru aprecierea fluenţei traficului. Prima abordare se bazează pe presupunerea că o staţie de bază cu trafic greoi are un risc redus de saturaţie. Prin urmare, este necesar să se aprecieze riscul de saturaţie a fiecărei staţii de bază. Acest lucru este echivalent cu estimarea momentului în care o staţie de bază se va satura. Deci, primul obiectiv al acestei teze este de a propune o abordare pentru predicţia seriilor de timp. Această abordare se bazează pe o descompunere multirezoluţie a semnalului cu ajutorul transformatei wavelet staţionare şi modele ARIMA. Aplicată la toate trasele din baza de date, această abordare a permis o primă clasificare a staţiilor de bază din punctul de vedere al fluenţei traficului, prezentată la sfârşitul Capitolului 3.

A doua abordare pentru aprecierea fluenţei traficului se bazează pe analiza dependenţei pe termen lung (LRD). Acesta este un concept statistic relativ nou în analiza traficului de comunicaţii şi poate fi implementat folosind undişoarele. Această proprietate are implicaţii importante privind performanţa, proiectarea şi dimensionarea reţelei. Estimarea gradului de dependenţă pe termen lung se realizează prin estimarea parametrului Hurst al seriei temporale analizate. Prin efectuarea de simulări şi analize, rezultatele noastre demonstrează că traficul din reţeaua WiMAX prezintă un comportament dependent pe termen lung. Obiectivul nostru în Capitolul 4 este de a evidenţia particularităţile traficului WiMAX din perspectiva dependenţei pe termen lung şi de a clasifica staţiile de bază pe baza fluenţei traficului. Această clasificare a staţiilor de bază este prezentată la sfâşitul Capitolului 4 şi este în acord cu clasificarea de la sfârşitul Capitolului 3, în ciuda faptului că ambele clasificări au fost efectuate prin estimări statistice. Din acest motiv, răspunsul la întrebarea pusă la începutul tezei este afirmativ, staţiile de bază prost poziţionate în topologia unei reţele WiMAX pot fi identificate prin analiza traficului. Rezultatele arată staţiile de bază care au o localizare bună în topologia reţelei făcând posibilă identificarea staţiilor de bază care au o localizare proastă. Acestea din urmă trebuie să fie repoziţionate în cadrul urmatoarei sesiuni de întreţinere a reţelei.

Capitolul 5 este dedicat concluyiilor şi contribuţiilor perosnale. Rezultatele acestei teze sunt atât de natură teoretică cât şi practică. Dintre rezultatele teoretice ar putea fi menţionate următoarele: analiza de ordinul al doilea a coeficienţilor wavelet prezentată în Capitolul 2, estimatorul parametrului Hurst bazat pe undişoare potrivit pentru serii de timp staţionare în sens larg propus în Capitolul 2, precum şi, un nou test de staţionaritate bazat pe reiterarea metodologiei Box-Jenkins propusă în Capitolul 3.

Printre rezultatele practice ale tezei pot fi menţionate următoarele. In Capitolul 3, am testat un algoritm de predictie pt serii de timp propus pentru retelele cu fir, in cazul retelelor wireless. Aceasta metoda se bazeaza pe transformata wavelet stationara (SWT) si tehnici statistice de analiza ale seriilor de timp. Contributiile principale sunt:

- Am validat utilitatea algoritmului propus in [Papagiannaki et al , 2003], in cazul de retele fara fir.

- S-au obtinut estimari exacte cu un cost minim de calcul. Toate estimarile noastre au fost obtinute în câteva secunde.

- Am identificat statiile de baza cu risc ridicat de saturatie.

- Am propus o strategie de selectare a undisoarelor mama, pe baza localizarii lor timp-frecventa. Am demonstrat prin simulari ca, in cazul de traficului de comunicatii, localizarea in timp este cea mai importanta caracteristica a undisoarelor mama utilizate pentru a calcula SWT.

- Am demonstrat ca cele mai bune rezultate sunt obtinute cu ajutorul undisoarei Haar. De asemenea, am demonstrat ca SWT reprezinta cea mai buna alegere intre transformatele in undisoare pentru predictia traficului fara fir. Acest lucru se datoreaza invariantei acesteia la translatii.

- Algoritmul este suficient de flexibil pentru a lucra cu seturi de date diferite, cum ar fi trafic de comunicatii, date financiare sau date de transport, fara a necesita modificari importante.

- Am comparat algoritmul de predictie propus cu alti algoritmi dezvoltati in echipa noastra de cercetare, bazati pe retele neuronale, si am dovedit utilitatea algoritmului nostru pentru predictii pe termen lung. Algoritmul propus de noi este mai rapid decat alti algoritmi ca urmare a utilizarii undisoarelor, ca urmare a utilizarii de MRA si ca urmare a utilizarii medierilor pe saptamana.

In Capitolul 4 am analizat traficul in cadrul retelei WiMAX cu scopul de a identifica particularitati ale retelei. Strategia aleasa pentru acest scop se bazeaza pe dependenta pe termen lung a traficului. Prezenta dependentei pe termen lung (LRD) in traficul de retea are un impact semnificativ asupra performantei retelei. Performanta unei retele de comunicatii fara fir depinde de o arhitectura eficienta (de pozitionarea buna a statiilor de baza). Contributiile principale pot fi rezumate dupa cum urmeaza:

- Am analizat traficul in uplink si downlink si am demonstrat ca traficul prezinta LRD.

- Am identificat o cauza pentru aparitia LRD care este tipica pentru retele fara fir: periodicitatile de o saptamana si de o zi.

- Utilizand estimatorului R/S pt parametrului Hurst, am dovedit ca dependenta pe termen lung poate fi redusa prin divizarea seriilor de timp corespunzatoare fiecarei statii de baza in serii cu durata de o zi. Am demonstrat ca in mod normal, traficul zilnic printr-o statie de baza nu ar trebui sa manifeste LRD.

- Am propus o metoda de estimare bazata pe undisoare care a dat o mai buna performanta in comparatie cu estimatorul R/S.

- Am comparat o serie de estimatori estimatorilor ale parametrului Hurst si am demonstrat prin simulari superioritatea estimatorului bazat pe undisoare.

- Am analizat pozitionarea statiilor de baza in arhitectura retelei WiMAX. Statiile de baza pentru care numarul de zile in care traficul este dependent pe termen lung atat in uplink cat si in downlink este ridicat sunt incorecte pozitionate.

- Rezultatele arata statiile de baza care au o localizare buna in topologia retelei si cele care nu sunt bine localizate. Acele statii de baza care nu au o localizare buna ar trebui sa fie repozitionate in viitor.

- Am observat ca statiile de baza care nu au o localizare buna, au un risc redus de saturatie. Aceasta observatie permite sa se verifice rezultatele estimarii din Capitolul 3, cu ajutorul rezultatelor prezentate în Capitolul 4 si vice-versa. Dintr-un numar total de 66 de statii de baza rezultatele analizei de pozitionare efectuate nu sunt concludente numai pentru o singura statie.

Ca şi perspective am putea enumera urmatoarele. Aparitia LRD ar putea fi rezultatul unor anomalii care apar în timpul anumitor zile. O metoda pentru a identifica anomalii de trafic ar fi foarte interesantă în continuarea cercetărilor. Un alt subiect consta în analiza statistica a coeficenților DWT de ordinul doi pentru procese nestaționare.