Využitie dátových úložísk v pracovnom prostredí MATLAB

Using the data storage in the working environment MATLAB

Stanislav Horal, Katedra aplikovanej informatiky, FPV UCM v Trnave

Michaela Horalová Kalinová, Katedra aplikovanej informatiky, FPV UCM v Trnave

Abstract: In term of system identification it is important to estimate the system module and to specify the system parameters. In that, the data for the identification process are pretty much the experimental data, it matters to deal with suitable data organization, as well as correspondent data storage. The suitable tool for automatic system identification is MATLAB, the environment for realization the scientist - research calculations. For integrated support working with the database systems we can provide wide data base for identification complex systems. The main goal this contribution is to point on possibilities of co-operation the MATLAB environment with database system providing the access to the relation databases and to the data storage. In this article we focus to the description the interface, communication ways, gathering and processing results.

Key words: system, identification, MATLAB, relational database, data warehouse, real time

Abstrakt: Z hľadiska identifikácie systémov má zásadný význam stanovenie modelu systému a špecifikovanie parametrov. Keďže údaje pre proces identifikácie sú do veľkej miery experimentálne údaje, má význam zaoberať vhodnou organizáciou údajov, ako aj príslušným dátovým úložiskom. Vhodným nástrojom pre automatickú identifikáciu systémov sa javí prostredie pre realizáciu vedecko-výskumných výpočtov MATLAB. Vďaka integrovanej podpore práce s databázovými systémami môžeme poskytnúť širokú dátovú základňu pre identifikáciu zložitých systémov. Cieľom príspevku je poukázať na možnosti spolupráce prostredia MATLAB s databázovým systémom poskytujúcim prístup k relačným databázam a dátovému skladu. V článku sa zameriame na popis rozhrania, spôsoby komunikácie, získanie a spracovanie výsledkov.

Kľúčové slová: systém, identifikácia, MATLAB, relačná databáza, dátový sklad, reálny čas

1. Úvod

Pojem identifikácie môžeme predstaviť ako výber modelu systému z určitej triedy systémov, ktorý svojou štruktúrou a vlastnosťami zodpovedá analyzovanému systému. Vhodný nástroj pre podporu identifikačného procesu predstavuje prostredie MATLAB. Prostredie pre vedecko-výskumné výpočty obsahuje integrovaný nástroj pre identifikáciu lineárnych aj nelineárnych dynamických systémov – System Identification Toolbox. Tento integrovaný nástroj umožňuje zastrešiť proces identifikácie od zberu a predspracovania údajov, cez odhad štruktúry systému až po simuláciu, predikciu výstupného signálu a riadenie systému.

Cieľom príspevku je poukázať na možnosti spolupráce prostredia MATLAB s databázovým systémom poskytujúcim prístup k relačným databázam a dátovému skladu.

2. Identifikácia v prostredí MATLAB

Pre proces identifikácie je v MATLAB-e dostupný modul System Identification Toolbox. Tento modul umožňuje odhadovať parametre lineárnych a nelineárnych modelov dynamických systémov z nameraných dát. Modul v svojom grafickom rozhraní nám umožní predpripraviť a realizovať proces identifikácie bez nutnosti poznať presnú syntax príkazov identifikačných metód a použitie ich parametrov.

Toolbox umožňuje importovať namerané dáta v časovej alebo frekvenčnej oblasti nachádzajúce sa v premenných prostredia MATLAB-u. Dáta môžu byť importované do prostredia z externých dátových súborov, mohli vzniknúť ako výsledok predchádzajúcich simulácií alebo mohli byť manuálne vytvorené v dynamických štruktúrach pracovného prostredia, napríklad prostredníctvom príkazového riadku. Pred samotnou identifikáciou je možné previesť predspracovanie dát prostredníctvom výberu transformácie dát, filtrov, výberom a zlučovaním existujúcich experimentov, resamplovaním, odstraňovaním trendov, výberom rozsahu alebo použitím ďalších nástrojov. To nám umožní dosiahnuť lepšie výsledky pri nasledujúcom procese identifikácie.

Proces identifikácie nám umožní odhadnúť parametre modelov dynamických systémov. Pri lineárnych modeloch máme možnosť použiť metódy pre identifikáciu modelov s frekvenčnou charakteristikou, identifikáciu modelov s impulznou charakteristikou, identifikáciu modelov s prenosom, identifikáciu vstupno-výstupných polynomiálnych modelov a identifikáciu stavovo-priestorových modelov. Pri nelineárnych modeloch môžeme použiť metódy pre identifikáciu nelineárnych ARX modelov a Hammerstein-Wienerove modely. Pri modeloch

identifikácie časových radov môžeme použiť metódu pre identifikáciu modelov s frekvenčnou charakteristikou.

3. Dátové pozadie objektov identifikácie

System Identification Toolbox poskytuje pre možnosť zjednodušenia práce s viacerými premennými dva špeciálne dátové objekty, ktoré v sebe zapuzdrujú hodnoty a vlastnosti dát. Dátové objekty iddata a idfrd sú najčastejšie napĺňané dátami z externých zdrojov alebo ich obsah vzniká na základe predchádzajúcich experimentov alebo simulácií.

Objekt iddata reprezentuje dáta v časovej alebo frekvenčnej oblasti. S výnimkou pri klonovaní objektu sa musí na vytvorení objektu podieľať premenná obsahujúca výstupný signál zo systému, ktorá je stĺpcovým vektorom alebo maticou, podľa násobnosti výstupného signálu systému. Vstupný signál je nepovinný, v prípade jeho uvedenia je reprezentovaný taktiež stĺpcovým vektorom alebo maticou, podľa násobnosti vstupného signálu systému. Vstupné aj výstupné dáta musia byť z tej istej oblasti. Na vytváraní objektu iddata sa ďalej podieľa časový interval po sebe nasledujúcich vzoriek dát, uvádzaný v sekundách. Objekt iddata môže byť taktiež vytvorený z objektu idfrd.

Najdôležitejšími spoločnými vlastnosťami objektu iddata, ktorý reprezentuje dáta v časovej alebo frekvenčnej oblasti, je špecifikácia oblasti, názov experimentu, názov premennej uchovávajúcej vstupný signál systému, názvy a jednotky jednotlivých vstupných kanálov, perióda vstupného signálu, názov premennej uchovávajúcej výstupný signál systému a názvy a jednotky jednotlivých výstupných kanálov.

V prípade, že objekt iddata reprezentuje dáta v časovej oblasti, vlastnosti objektu obsahujú prepočítané hodnoty časového vektora, informácie o časovej jednotke a počiatočnú hodnotu časového vektora. Naopak, v prípade, že objekt iddata reprezentuje dáta vo frekvenčnej oblasti, vo vlastnostiach objektu sú uchovávané hodnoty frekvencií pre definovanie Fourierovej transformácie signálov a jednotku frekvencie.

Objekt idfrd reprezentuje dáta alebo model s frekvenčnou charakteristikou. Pri vytváraní objektu je potrebné poznať frekvenčnú charakteristiku lineárneho systému s frekvenčnými hodnotami a intervalom vzorkovania pre diskrétne systémy. Frekvenčná charakteristika je reprezentovaná trojrozmerným poľom, kde jednotlivými dimenziami je počet výstupov, počet vstupov a počet frekvencií. Frekvencia je reprezentovaná stĺpcovým vektorom.

Najdôležitejšími vlastnosťami objektu idfrd je frekvenčná charakteristika, frekvencia, vektor frekvenčných, vstupných a výstupných jednotiek.

4. Relačná databáza ako úložisko pre experimentálnu identifikáciu

Pri identifikácii zložitých technologických systémov prevažne vychádzame z hodnôt meraných fyzikálnych veličín. Takúto širokú údajovú základňu, je potrebné reprezentovať prostredníctvom adekvátnych dátových štruktúr a uchovávať vo vhodných dátových úložiskách. Najjednoduchším spôsobom, akým možno uchovávať a spracovávať veľké množstvo dát je použitie relačných databáz.

Relačné databázy predstavujú databázový systém založený na relačnom modeli dát a relačnej algebre. Relačný databázový model je tvorený sústavou normalizovaných, v čase meniacich sa databázových tabuliek.

Výstupný signál dynamického systému môžeme vo všeobecnosti považovať za viacnásobný (viackanálový) výstupný signál. Hodnoty výstupného signálu môžeme vkladať do samostatnej databázovej tabuľky tak, aby sme mali k dispozícii maticu n x m, kde n reprezentuje počet hodnôt výstupného signálu a m počet výstupných veličín (počet kanálov). Atribútmi takejto databázovej tabuľky môžu byť napr. identifikátor dynamického systému, časová známka alebo číselné označenie poradia stavu dynamického systému, poradové číslo výstupnej veličiny alebo jej identifikátor, poradové číslo hodnoty a samotná hodnota výstupného signálu dynamického systému.

Obdobne, vstupný signál dynamického systému môžeme vo všeobecnosti považovať za viacnásobný vstupný signál. Rovnakým spôsobom teda môžeme do databázovej tabuľky uložiť hodnoty vstupného signálu. Hodnoty vstupných veličín vstupného signálu a hodnoty výstupných veličín výstupného signálu môžeme ukladať v podobe reálnych čísel s požadovanou presnosťou podľa aplikačných požiadaviek, alebo v podobe celých čísel pri tolerovaní určitej odchýlky. Vo všeobecnosti informačný systém nemá obmedzenie šírky dát, na druhej strane nás však zaujíma rýchlosť spracovania dát. Keďže čas spracovania dát je minimálny pri minimálnej šírke dát, je potrebné určiť minimálnu šírku dát z hľadiska metrologických charakteristík merania dát. Metrologická charakteristika snímania dát môže byť základom určenia entropie informácie a veľkosť informačného prúdu. V praxi nemá zmysel ukladať dáta s väčšou šírkou dát ako sme určili z informačného prúdu.

Ďalšou možnosťou je v databázových tabuľkách uchovávať serializované hodnoty jednej vstupnej alebo výstupnej veličiny v jednom konkrétnom stave dynamického systému, čo sa však prieči normálovým formám a taktiež zavádza ďalšiu réžiu pri spracovávaní dát, napr. pri procese extrakcie dát, ich transformácie a zavádzania do dátového skladu (ETL).

Ďalšie doplňujúce informácie potrebné pre vytvorenie objektu iddata budeme ukladať do tabuľky so všeobecnými informáciami o skúmanom dynamickom systéme. Okrem jeho charakteristiky v prirodzenom jazyku musíme mať k dispozícii hodnotu intervalu medzi po sebe nasledujúcimi vzorkami dát. Táto hodnota je nulová pre dáta vo frekvenčnej oblasti v spojitom čase.

Pre prístup k externým dátam z databázových systémov môžeme použiť nástroj MATLAB-u Database Toolbox. Tento nástroj vo všeobecnosti dokáže prevádzať import a export dát medzi prostredím MATLAB-u a takmer ľubovoľným databázovým systémom. Použitým komunikačným rozhraním môže byť ODBC (Open Database Connectivity) alebo JDBC (Java Database Connectivity) podľa podpory rozhrania používaným operačným systémom. Tým máme zabezpečenú podporu od Excel-u a Access-u cez menšie databázové systémy ako sú MySQL a PostgreSQL až po komplexné databázové riešenia ako sú napr. MSSQL, Informix a Oracle.

Grafické rozhranie tohto nástroja umožní vytvoriť SQL dotaz výberom konkrétneho dátového zdroja, katalógu, schémy, databázových tabuliek a polí. Pri výbere je možné vizuálne previesť definíciu reštrikcie, zgrupovania, dodatočných podmienok a triedenia. Pri vytvorení SQL dotazu alebo jeho manuálnych vložením je výsledok dotazu vložený do definovanej premennej prostredia MATLAB-u. Dátový typ tejto premennej predstavuje maticu n x m, kde n je počet získaných záznamov a m počet stĺpcov dotazu. Následne je potrebné previesť transformáciu takto získaného výsledku na stĺpcový vektor alebo na maticu, podľa násobnosti vstupného alebo výstupného signálu systému.

5. Použitie dátových skladov reálneho času

Dátový sklad je podnikovo štruktúrované úložisko subjektovo orientovaných, integrovaných, časovo premenlivých, historických dát použitých na získavanie informácií a podporu rozhodovania. V dátovom sklade sú uložené atomárne a sumárne dáta.

Dáta prítomné v dátovom sklade sú integrované z viacerých typicky nehomogénnych zdrojov, pričom v procese integrácie je zabezpečená transformácia údajov do jednotného výsledného

formátu, reštrukturalizácia dát do adekvátne nadefinovaných štruktúr, ako aj kontrola kvality údajov. Údaje v dátovom sklade reprezentujú stavy a hodnoty reálnych procesov v konkrétnom časovom okamihu.

Keď sú dáta vkladané do dátového skladu reálneho času, dáta sú vkladané v reálnom čase, nie je možné plánovať neustále technické prestávky pre vkladanie dát a zároveň je nutné dáta ponechávať v konzistentnom stave. V súčasnosti so vzrastajúcou požiadavkou na riešenie niektorých problémov a analýz v reálnom čase sa na trhu objavujú pokročilé ETL nástroje zabezpečujúce prevedenie korektného procesu ETL v reálnom čase. Riešenie, keď ETL proces neprebieha v reálnom čase ale len v takmer reálnom čase, je použiteľné pre dynamické systémy, kde proces ETL a následná analýza sú minimálne dva krát rýchlejšie ako frekvencia riadenia dynamického systému. Ak by táto požiadavka nebola splnená v celom pásme frekvencie riadenia, je možné na riadenie dynamického systému využiť len nižšie frekvencie a vyššie frekvencie budú využité na predikciu.

Výhodou použitia dátových skladov reálneho času je nielen prispôsobené úložisko pre ukladanie veľkého množstva historických dát, ale dostupnosť analytických služieb a nástrojov, ktoré dátové sklady majú k dispozícií. OLAP (Online Analytical Processing) nástroje vykonávajú analytické a reportovacie činnosti, ktoré sú využiteľné pri identifikácii dynamických systémov. (4, 5)

Data mining je proces analýzy dát z rôznych perspektív a ich premena na užitočné informácie. Z matematického a štatistického hľadiska ide o hľadanie korelácií, teda vzájomných vzťahov alebo vzorov v dátach. Týmto spôsobom môžeme objaviť nové závislosti medzi meranými veličinami, ktoré by boli ťažko odhalené klasickými metódami identifikácie.

Data miningové časti analytických služieb sú s každou novou verziou doplňované o ďalšie algoritmy. V SQL Serveri 2008 je v súčasnosti dostupných 12 algoritmov, v Oracle Data Mining je v súčasnosti dostupných taktiež 12 data mining algoritmov.

Väčšina dodávateľov analytických nástrojov obsahuje nástroje s grafickým rozhraním pre vizuálny návrh predikčného dotazu. Typicky je možné pomocou sprievodcu vybrať data miningový model, použitý algoritmus, zdroj dát, databázové stĺpce, na ktorých závisí výsledok analýzy, databázové stĺpce, ktorých hodnoty chceme predikovať a ďalšie informácie. Výsledkom vizuálneho návrhu je zložitý, ale komplexný predikčný dotaz. Podobne, ako príkazy na vytvorenie OLAP kocky, napĺňanie tabuliek faktov a dimenzií dátami, aj predikčný dotaz je opísaný jazykom SQL, ktorý môžeme použiť v akejkoľvek

databázovej aplikácii. (2, 3) Pre naše účely výsledky predikčného dotazu môžeme spracovať prostredím MATLAB-u.

6. Zhrnutie

V príspevku sme sa snažili ukázať možnosti spolupráce prostredia pre vedecko-výskumné výpočty MATLAB s databázovým systémami poskytujúcim prístup k relačným databázam a dátovému skladu. Predstavili sme modul pre identifikáciu dynamických systémov, jeho dva základné dátové objekty a modul pre prácu s databázovými systémami. Ukázali sme možnosť využitia relačných databáz ako dátového úložiska pre proces experimentálnej identifikácie, navrhli sme spôsob ukladania vstupných a výstupných veličín a ukázali sme, akým spôsobom je možné informácie vniesť do prostredia MATLAB. V príspevku sme ďalej ukázali možnosť využitia dátových skladov a analytických služieb.

7. Zoznam bibliografických odkazov

(1) Hudzovič, P.: Identifikácia a modelovanie. Bratislava, ES SVŠT, 1986.

(2) Lacko, L.: Business Intelligence v SQL Serveru 2005. Computer Press, Brno, 2006. ISBN 80-251-1110-5.

(3) Kimball, R., Caserta, J. 2004. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley Publishing, Inc., Indianapolis, 2004. ISBN: 978-0-7645-6757-5.

(4) Bruckner, R. M., Jeng, J. J., Schiefer, J.: Real-time Workflow Audit Data Integration into Data Warehouse System. ECIS, Naples, 2003.
On-line [http://en.scientificcommons.org/50066073], [cit.:15.4.2010]

(5) Horal, S., Kalinová M., Michaľčonok, G. F.: Real-time data warehouse. Infokommunikacionnye technologii v nauke, proizvodstve i obrazovanii: tretja meždunarodnaja naučno-trechničeskaja konferencija. Stavropoľ, 2008. UDK 004 (06).

8. Adresy autorov:

Stanislav Horal, Mgr.
Katedra aplikovanej informatiky FPV UCM
Nám. J. Herdu 2
917 01 Trnava
stanislav.horal@ucm.sk

Michaela Horalová Kalinová, Mgr.
Katedra aplikovanej informatiky FPV UCM
Nám. J. Herdu 2
917 01 Trnava
michaela.horalova@ucm.sk