XI. FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA

Kolozsvár, 2006. március 24-25.

A PREDIKTÍV HATÉKONYSÁG NÖVELÉSE REGRESSZIÓS FA ALGORITMUSSAL

Fehér Tamás

Abstract. Tasks, where the value of an unknown parameter has to be estimated, are typical in data mining projects. The solution of this kind of problems requires the creating of a predictive model. There are several methods for the solution of this type of tasks, e. g. the decision trees and the regression. In the paper an algorithm is demonstrated which combines the properties of the two techniques. Tests on different datasets show the efficiency of the method compared to the results which were provided by the traditional regression and decision tree algorithms.

Összefoglaló. Az adatbányászati projektekben tipikusak az olyan feladatok, amelyben egy ismeretlen paraméter értékét kell becsülni. Azt a folyamatot, amely során elkészül a modell, melynek segítségével eddig nem ismert adatok alapján kell becslést tenni az ismeretlen paraméterre, prediktív modellezésnek nevezik. Számos módszer létezik az ilyen típusú feladatok megoldására. A döntési fák és a regresszió is ilyen módszerek. A cikkben bemutatásra kerül egy olyan algoritmus, amely a két technika előnyeit egyesíti. Több adathalmazon elvégzett futtatás is megmutatja a módszer hatékonyságát, összehasonlításban a hagyományos regresszió és döntési fa algoritmusok által szolgáltatott eredményekkel.

1. Bevezetés

A hagyományos döntési fák alapvetően klasszifikációs feladatok megoldására szolgálnak. Minden egyes levélben egy osztálycímke található, amely megadja, hogy a kérdéses objektum melyik osztályba esik. A CART [1] algoritmus lehetővé teszi folytonos változók becslését döntési fa segítségével. Az algoritmus által felépített fák leveleibe egy konstans függvény kerül: ez ad becslést az adott paraméter értékére. Az évek folyamán további algoritmusok kerültek kifejlesztésre, amelyek a folytonos érték hatékonyabb becsléséhez lineáris függvényeket helyeznek a levelekbe (M5 [2], RETIS [3], TSIR [5], SMOTI [6]).

Az algoritmusok közti különbség a fa felépítésében van. A legfontosabb tényező, hogy hogyan találja meg az algoritmus a vágási pontokat. A legtöbb algoritmus valamilyen szórás alapú megközelítést használ a vágás jóségának mértékeként.

2. Regressziós modellek a levelekben

Abban az esetben, ha lineáris regressziót szeretnénk a fa leveleibe tenni, nem feltétlenül célravezető, ha a szórás alapján ítéljük meg a vágás jóságát [4]. Az 1. ábrát megfigyelve látszik: bár a szórás a két csoportban az 'a' pontban történő vágás mentén lenne minimális, mégis célszerűnek látszik egy újfajta

1. ábra

jósági mérték bevezetése, és a vágást a 'b' pontban végrehajtani. Ez alapján a vágás jóságának mértéke a vágás bal és jobb oldalára illeszthető regressziós modell illeszkedési hibája lehet, melyet a következő formula definiál:

$$I(E) = \frac{1}{N_l + N_r} \left[\frac{1}{N_l} \sum_{e_i \in E_l} (y_i - g_l(\bar{x}_i))^2 + \frac{1}{N_r} \sum_{e_j \in E_r} (y_j - g_r(\bar{x}_j))^2 \right]. \quad (1)$$

$I(E)$ jelöli a vágás jóságának mértékét. N_l és N_r a vágás bal és jobb oldalára eső elemek száma, E_l és E_r a bal és jobb oldal elemeinek halmaza. A g_l és g_r függvény a bal-, illetve jobboldali gyermekcsomópontok elemeire illesztett regressziós hiperfelületet reprezentálja.

A módszer hátránya a óriási számításigény. Az eredeti cikkben [4] a szerző csupán néhány tíz megfigyelést tartalmazó adathalmazokon próbálta ki az algoritmust. Bizonyos kompromisszumok árán azonban ennél lényegesen nagyobb mennyiségű adatot tartalmazó adathalmazon is alkalmazható a módszer.

Tekintve, hogy a vágás jóságának mértéke a keletkező két csoportra illeszthető regressziós modell illeszkedési hibája, elviekben végig kellene próbálni az összes létező vágást az összes dimenzió mentén ahhoz, hogy el lehessen dönteni, az adott szinten mi a legmegfelelőbb vágás. Ez egy nagy dimenziószámú, illetve nagy elemszámú adathalmazon túl nagy számítási igényt generál.

Ahhoz, hogy ésszerű időn belül futási eredményeket kapjunk, kompromisszumot kell kötni. Ha nem végezzük el az összes létező vágást, lényegesen le lehet faragni a futási időt. Az adatfelhő elhelyezkedéséről, formájáról mintavételezéssel is információt szerezhetünk, vagyis a dimenziómenti vágásokat csak bizonyos helyeken elvégezve is egy jó megoldáshoz juthatunk.

3. A regressziós fa algoritmus

A módszer magját egy olyan eljárás képezi, amely a bemenetként megkapott csomóponton az összes vágást elvégzi (egy külön algoritmus jelöli ki a nem túl nagy számú vágási pontot), a vágások közül pedig kiválasztja azt, ami mentén elvágva az adathalmazt a legkisebb hibával illeszthető a két regressziós hiperfelület (az átlagos hiba, így a vágás jóságának mértéke a bal és jobb oldali elemszámokkal súlyozott átlagos négyzetes eltérések számtani közepe). Ehhez minden egyes vágásnál elvégzi a szükséges regressziókat.

A fa felépítése során ez az eljárás hajtódik végre iteratív módon minden csomópontra, megadott számú szinten. Az iterációk száma minden szinten megegyezik az előző szint csomópontjainak a számával. Minden iterációs lépésben a korábbi szint egy csomópontja fogja az eljárás inputját képezni.

Az algoritmus SAS-ban [8] kódolva készült el. A regressziós lépésben a program egy stepwise regressziót hajt végre minden esetben.

A dimenziónkénti vágási pontok meghatározására egy külön algoritmus került kifejlesztésre, amelynek részleteit helyhiány miatt ebben a cikkben nincs mód bemutatni.

Az algoritmus kimenete egy tábla, amely minden szükséges információt tartalmaz a fa felépítéséhez. Ezt szemlélteti a 2. ábra.

ID	Leaf	Rule	N	Intercept	Param1	Param2	...
1	0	attr1 ≤ 25	4000	123.456	0.789	2.345	...
2	0	attr1 > 25	6000	234.567	1.234	5.432	...
3	1	attr1 ≤ 25 & attr2 ≤ 100	1000	87.654	0	0.987	...
4	1	attr1 ≤ 25 & attr2 > 100	3000	34.567	0.456	0.023	...
5	1	attr1 > 25 & attr3 ≤ 60	4000	54.321	2.345	0	...
6	1	attr1 > 25 & attr3 > 60	2000	12.345	0.456	0.789	...

2. ábra

4. Kísérleti eredmények

Három különböző adathalmazon történő futtatásból származnak az eredmények (1. táblázat). Mindhárom adathalmaz esetén először egy sima regressziós modell, majd a SAS által implementált döntési fa algoritmus, végül a regressziós fa algoritmus becsli a célváltozó értékét. Az algoritmusok

prediktív teljesítményének mértéke a teszttáblományok elemeinek a becstelt értékektől vett átlagos relatív eltérése. Tesztelési célra mindhárom esetben az adatok 10%-a lett elkülönítve.

1. táblázat

Adatok	Alkalmazott módszer	Átlagos relatív hiba (%)
- Mesterséges adathalmaz - 100 rekord / 3 dimenzió - Folytonos célváltozó, 1 bináris és 1 folytonos attribútum	Regresszió	120.92
	Döntési fa	104.43
	Regressziós fa	5.18
- Orvosi adatbázis - 470 rekord / 10 dimenzió - Folytonos attribútumok	Regresszió	2.84
	Döntési fa	4.59
	Regressziós fa	0.52
- Banki adatbázis - 20018 rekord / 10 dimenzió - Folytonos attribútumok	Regresszió	83.10
	Döntési fa	53.08
	Regressziós fa	46.76

5. Összefoglalás

A teszteredmények fényében megállapítható, hogy a cikkben prezentált algoritmus prediktív pontosság tekintetében jobb, mint a hagyományos regresszió, de a SAS által implementált döntési fa algoritmusnál is pontosabb becslésekkel szolgál. A teljesítmény ára a viszonylag magas futási idő, amely 20 perc körül adódik egy átlagos gépen (PIII, 850 MHz, 384 MB RAM), egy nagyméretű, ötszintű fa esetén (~20000 rekord, 10 attribútum).

Irodalomjegyzék

- [1] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE J.: *Classification and regression tree*, Wadsworth & Brooks, 1984.
- [2] QUINLAN J. R.: *Learning with continuous classes*, in Proceedings AI'92, Adams & Sterling (Eds.), World Scientific, pp. 343-348, 1992.
- [3] LUBINSKY D.: *Tree Structured Interpretable Regression*, in Learning from Data, Fisher D. & Lenz H.J. (Eds.), Lecture Notes in Statistics, 112, Springer, pp. 387-398, 1994.
- [4] KARALIC, A.: *Linear Regression in Regression Tree Leaves*. In Proceedings of ISSEK'92 (International School for Synthesis of Expert Knowledge) Workshop, Bled, Slovenia, 1992.
- [5] MALERBA, D., APPICE, A., CECI, M., MONOPOLI, M.: *Trading-off local versus global effects of regression nodes in model trees*. In H.-S. Hacid, Z.W. Ras, D.A. Zighed & Y. Kodratoff (Eds.), Foundations of Intelligent Systems, 13th International Symposium, ISMIS'2002, Lecture Notes in Artificial Intelligence, 2366, 393-402, Springer, Berlin, Germany, 2002.
- [6] TORGO, L.: *Functional Models for Regression Tree Leaves*. Proc. 14th International Conference on Machine Learning, 1997.
- [7] FAJSZI, B., CSER, L.: *Üzleti tudás az adatok mélyén*. Budapest, 2004.
- [8] SAS INSTITUTE INC.: <http://www.sas.com/>

Fehér Tamás, doktorandusz

Miskolci Egyetem, Alkalmazott Informatikai Tanszék (H-3515 Miskolc, Egyetemváros)

Tel.: +36 46/565-111/19-52, E-mail: feher@ait.iit.uni-miskolc.hu