CLASSIFICATION TELETRAFFIC SYSTEMS EXPRESSED THROUGH LOSS DECISION TREE

Georgiev Georgi, Balabanova Ivelina, Technical university of Gabrovo, givanow@abv.bg, ivstoeva@abv.bg

Abstract: Decision tree is a method defining rules in a hierarchical structure in the form of a coherent set of logical structures "if-then", on which is formed classification model derived from multiple instructional containing objects (observations) and their attributes. The dependent variables can be numeric, categorical and discrete. Most often these tasks are solved for binary classification (dichotomous classification model). All other classification problems and answers trees entering the internal nodes can be more than two.

Keywords: decision tree, classification method, K-fold, teletraffic, Artificial Neural Networks

КЛАСИФИЦИРАНЕ НА ТЕЛЕТРАФИЧНИ СИСТЕМИ С ЯВНИ ЗАГУБИ ПОСРЕДСТВОМ ДЪРВО НА РЕШЕНИЯТА

Георги Георгиев, Ивелина Балабанова, Технически университет - Габрово, givanow@abv.bg, ivstoeva@abv.bg

Абстракт: Дървото на решенията (Decision tree) е метод, дефиниращ правила в йерархична структура под формата на последователен набор от логически конструкции „if-then”, въз основа на които се формира класификационен модел, получен от обучаващо множество, съдържащо обекти (наблюденията) и техните атрибути (признаци). Зависимите променливи могат да бъдат числови, категориални и дискретни.

Ключови думи: дърво на решенията, класификационнен метод, телетрафик, изкуствени невронни мрежи

Най-често решаваните задачи са тези за бинарна класификация (дихотомен класификационен модел). При всички останали класификационни задачи отговорите и клоните на дърветата, влизащи във вътрешните възли могат да бъдат повече от два.

Предимства на класификационния модел - Дърво на решенията:

- Интуитивно разбиране на класификационния модел;
- Задаване на произволно количество променливи на входа на дървото. Алгоритъмът сам определя значимостта на променливите и избира тези, които ще участват при построяване на дървото. По този начин отпада необходимостта потребителя сам да избира входните променливи;

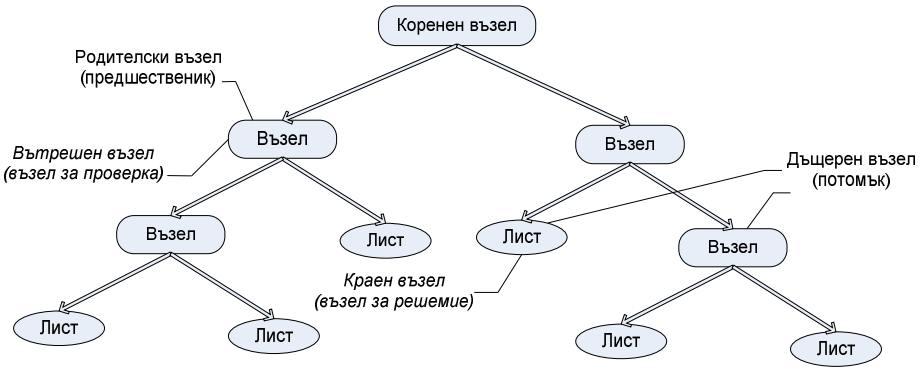
- Възможност за работа с числови и текстови данни;
- Алгоритмите за построяване на дървото работят и при липса на част от входните данни, извличане на правила на естествен език и др.

Недостатъци:

- Ограничение само до един изход;
- Нестабилност на генерираните решения в някои случаи;
- Дървовидните структури, създадени на база на числови масиви от данни, понякога могат да бъдат доста сложни при вземане на решения и др.

Приложение:

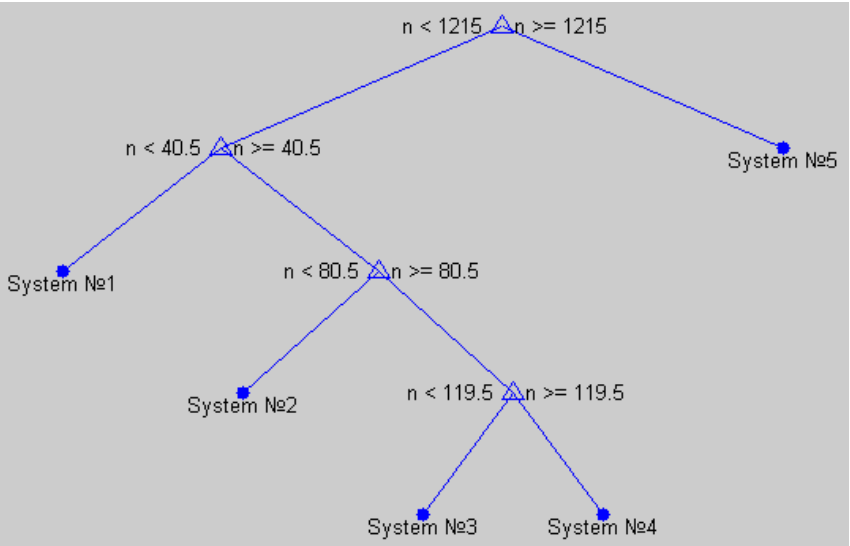
- Банково дело - при оценка на кредитоспособността на клиентите;
- Промишленост – контрол на качеството на произведената продукция и проверка на качеството на заваряване;
- Медицина – диагностициране на различни заболявания;
- Молекулярна биология – анализ на строежа на аминокиселини



Фиг.1. Класификационен модел дърво на решенията

На фиг.2. е представен класификационен модел, построен по метода дърво на решението, за определяне на груповата принадлежност на телетрафични системи с явни загуби при фиксирана вероятност за загуби B = 0.05. При създаване на класификатора е използван набор от двеста наблюдения, формирани от два информативни признака – постъпващият трафик A, изчислен по формулата на Ерланг, и информационните канали n. Целевите променливи са дефинирани посредством категориален тип данни. Обособени са пет класа телетрафични системи с различен диапазон на информационните канали, съответно при n = 1 до 40, n = 41 до 80, n = 81 до 120, n = 30 до 1200 и n = 1230 до 2400. Построеният класификационен модел е базиран на вторият

информативен признак (n), определен от алгоритъма като независима предсказваща променлива с по-голяма значимост.



Фиг. 2. Класификационен модел за определяне на принадлежността на телетрафични системи с явни загуби

Оценка на качеството на класификатора

При оценка на качеството на класификатора са използвани техническите подходи ресубституция и крос-валидиране.

1) Ресубституция. Оценката от ресубституция представлява частта на некоректни класификации, която се получава при прогнозиране на данни, участващи в процеса на обучение.

2) K-fold тип крос-валидиране. При приложение на указания подход входният набор от данни, съдържащ N наблюдения (еталона), се разделя на k подмножества като k-1-то се използва за обучение, а k-тото за тестване на крос-валидиращите модели. Процесът се повтаря k пъти. К е цяло число обикновено равно на 10, но по принцип остава нефиксиран параметър. При стойност на параметъра k = 10, следва че 10% от множеството от N наблюдения (еталона) ще бъде използвано за тестване, а 90% за обучение на крос-валидиращите модели. Оценката от крос-валидиране се определя като средното количество некоректни класификации на всеки крос-валидиращ модел при прогнозиране на данни, които не са използвани за обучение.

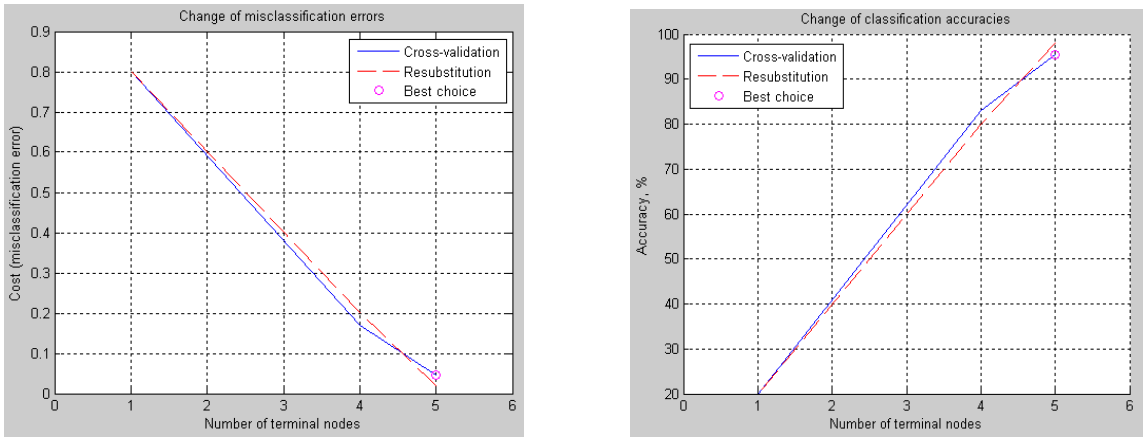
Получената класификационна точност при ресубституция често е твърде оптимистична при определяне на принадлежността на нови данни, ето защо като по-добра доверителна мярка се препоръчва точността от крос-валидиране. Средната точност между прилаганите технически подходи се приема като приблизително очаквана при класифициране на нови данни.

Резултати при оценка на качеството на класификатора

Извършена е оценка на качеството на класификатора с приложение на указаните технически подходи при различни нива на отсичане на разклонения от първоначално създадения модел. Търсено е оптимално ниво, при което се постигат най-ниска част на некоректни класификации от 0 до 100% (мащабирана в интервала от 0 до 1) и най-висока класификационна точност. При процеса крос-валидиране входният набор от данни в процентно съотношение е разделен на 75% за обучение и 25% за тестване на валидиращите модели. Получените резултати при оценка на качеството на класификатора са показани в табличен и графичен вид, съответно в табл. 1 и на фиг. 3.

Ниво на отсичане на разклонения	Брой възли	Ресубституция		Крос-валидиране	
		Част на некоректни класификации	Точност, %	Част на некоректни класификации	Точност, %
0	5	0.0200	98.00	0.0450	95.50
1	4	0.2000	80.00	0.1700	83.00
2	1	0.8000	20.00	0.8000	20.00

Таблица. 1 Резултати при оценка на качеството



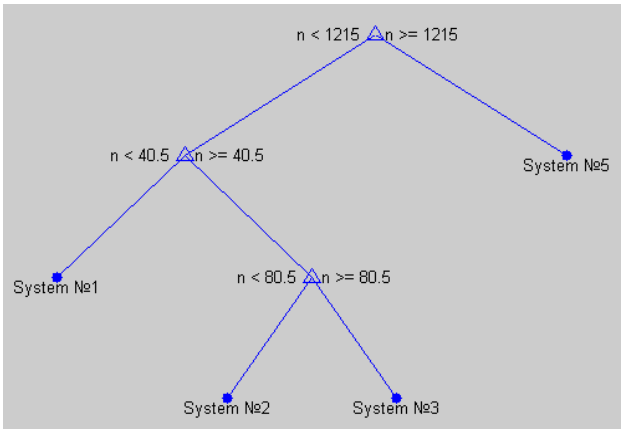
а) *б)*

Фиг. 3. Изменение на частите на некоректни класификации а) и класификационните точности б) при ресубституция и крос-валидиране

Най-високи класификационни точности при ресубституция от 98.00% и крос-валидиране от 95.50% са постигнати при нулево ниво на отсичане на разклонения. Установено е, че намереният оптимален е идентичен с първоначално генерираният

класификационен модел. Приблизително очакваните точности при определяне на принадлежността на нови наблюдения при нулево, първо и второ ниво, съответно са 96.75%, 81.50% и 20.00%.

Получените класификационни модели при първо и второ нива на отсичане на разклонения са представени на фиг. 4 и фиг. 5.



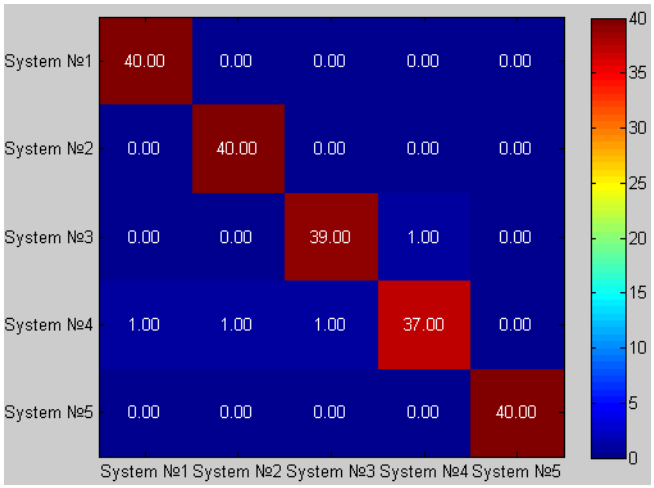
Фиг. 4. Класификационен модел при ниво на отсичане на разклонения 1

```
>> treelev2 = prune(tree,'Level',2)

treelev2 =

Decision tree for classification
1  class = System №1
```

Фиг. 5. Matlab изглед на класификационен модел при ниво на отсичане на разклонения 2

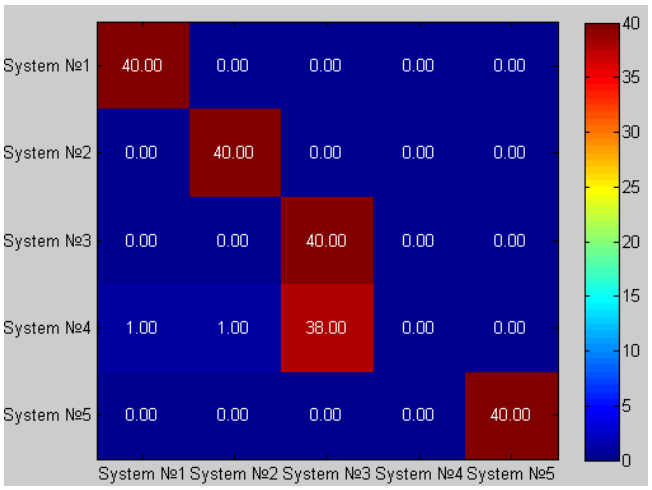
Класификационни матрици на наблюденията

Матриците на некоректни и коректни класификации при нулево, първо и второ ниво на отсичане на разклонения, показани на фиг. 6, фиг. 7 и фиг.8, илюстрират разпределението на наблюденията от състава на входния набор от данни по класове. Ако класификацията е протекла коректно, наблюденията трябва да бъдат разположени по диагонал на матриците, а всички останали техни елементи да са нули.



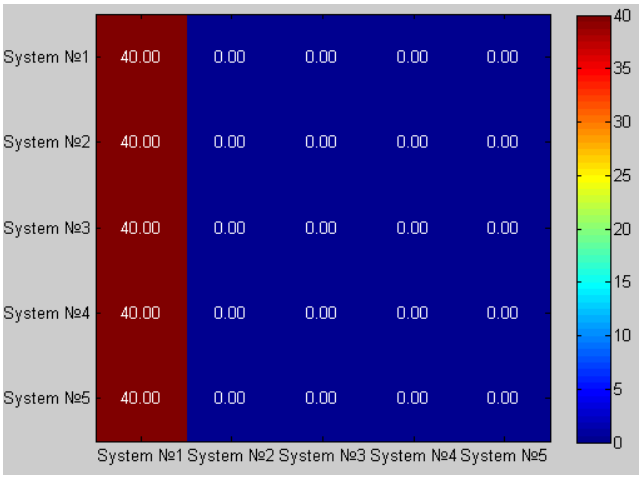
Фиг. 6. Класификационна матрица при ниво на отсичане на разклонения 0

Коректно класифициране на наблюденията е постигнато при телетрафични системи №1, №2 и №5. Едно наблюдение от система №3 е отнесено към №4, а три от №4 са класифицирани съответно към системи №1, №2 и №3.



Фиг. 7. Класификационна матрица при ниво на отсичане на разклонения 1

Некоректно класифицирани са всички наблюдения от телетрафична система №4 - две са с определена принадлежност към системи №1 и №2, а тридесет и осем са отнесени към №3. За останалите системи няма наличие на неправилни класификации.



Фиг. 8. Класификационна матрица при ниво на отсичане на разклонения 2

Всички наблюдения, с изключение на принадлежащите към телетрафична система №1, са некоректно отнесени към нея.

Сравнителен анализ между изкуствени невронни мрежи и дърво на решенията

Проведен е сравнителен анализ между апарата на изкуствените невронни мрежи и методът дърво на решенията по следните показатели:

-точност при класифициране;

-обем на използваната памет;

-време за обучение;

-време за класифициране.

Установено, е че при използване на невронните мрежи се постига висока, а при дървото на решенията средна класификационна точност. Малък е заеманият обем на използвана памет като хардуерен ресурс от обектите, създадени и по двата метода. Невронните мрежи изискват повече време за обучение за разлика от класификационните модели, построени по метода дърво на решенията. Двата апарата се характеризират с високо бързодействие при класифициране на данни. Те могат успешно да се прилагат в обучението на студентите по Телекомуникации.[1]

Във връзка с последният показател е извършено тестване на предложените архитектура на изкуствена невронна мрежа и класификатор по метода дърво на решенията за определяне на принадлежността на телетрафични системи с явни загуби с едни и същи тестови еталони. Получените резултати са представени в таблица 2. При дървото на решенията се забелязва значително по-високо бързодействие при класифициране на тестовите еталони в сравнение с изкуствената невронна мрежа. Това дава основание да се предполага, че подобна тенденция би се запазила при класифициране на нови данни.

Получени класификационни времена и резултати при тестване

Телетрафична система №1	Телетрафична система №2	Телетрафична система №3	Телетрафична система №4	Телетрафична система №5
Тестови еталони				
n = 23; A = 18.08	n = 59; A = 53.559	n = 101; A = 96.265	n = 480; A = 490.81	n = 1860; A = 1940.4
Изкуствена невронна мрежа				
Време, секунди	Време, секунди	Време, секунди	Време, секунди	Време, секунди
0.015574	0.018175	0.017045	0.020313	0.020835
Калкулиран резултат	Калкулиран резултат	Калкулиран резултат	Калкулиран резултат	Калкулиран резултат
1.1101 -0.1335 -0.0091 0.0329 0.0012	-0.0967 1.0639 -0.0371 0.0691 -0.0009	-0.0183 -0.0041 1.0755 -0.0532 0.0002	0.0067 0.0008 -0.0138 1.0004 -0.0027	-0.0001 -0.0000 0.0004 0.0003 0.9995
Дърво на решенията				
Време, секунди	Време, секунди	Време, секунди	Време, секунди	Време, секунди
0.002608	0.002762	0.003537	0.001787	0.002479
Предсказан клас	Предсказан клас	Предсказан клас	Предсказан клас	Предсказан клас
System №1	System №2	System №3	System №4	System №5

Таблица 2. Получени класификационни времена и резултати при тестване

Заключение

Предложеният модел, създаден по метода дърво на решението, успешно може да бъде използван при решаване на класификационни задачи за определяне на

принадлежността на телетрафични системи с явни загуби. Приложените технически подходи за оценка на качеството на класификатора показват високи точности от 98.00% при ресубституция и 95.50% при крос-валидиране. Базирайки се на правилото, че точността при крос-валидиране се взема с по-голяма достоверност, при класифициране на нови данни е приета приблизително очаквана точност от 96.75%. При прогнозиране на груповата принадлежност на тестови еталони, посредством селектираната структурата на невронна мрежа и построеният класификационен модел по метода дърво на решението, е постигнато по-голямо бързодействие при вторият подход. Приложението на апарата на изкуствените невронни мрежи при класифициране на телетрафични системи с явни загуби гарантира по-висока точност, но малко по-малко бързодействие в сравнение с дървото на решенията. Ето защо изборът на метод е компромисен вариант между търсено съотношение точност и бързодействие при класифициране на данни.

References:

[1] Димова Р, М. Иванов, В. Маркова, С. Костадинова, Интердисциплинарен подход за подобряване качеството на обучение по телекомуникации, Int. conf.UNITECH2013, , Габрово, България, 22-23.11.2013.

[2] Мирчев М. Телетрафично проектиране, издателство “Нови Знания”, София, 2002.

[3] Радев Д., Т. Илиев, Г. Христов. Компютърно моделиране на телетрафични системи., Изд. РУ “Ангел Кънчев”, Русе, 2008.

[4] Радев Д. Теория на телетрафика., Изд. “Нови Знания”, София, 2004.