Курсов проект
по
Извличане на знания от данни
(Data mining)

Poker rule induction

Изготвили:
Емил Гоцев, 61455
Валентин Змийчаров, 61481

Факултет по математика и информатика
4-ти курс Софтуерно инженерство
зимен семестър 2014/2015

Съдържание

Декларация за липса на плагиатство

Тази курсова работа е моя работа, като всички изречения, илюстрации и програми от други хора са изрично цитирани.

Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.

Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

Трите имена и подпис на студентите:

Валентин Венков Змийчаров

VZmiycharov

Емил Иванов Гоцев

EGotsev

1. Мотивация, Задача на курсовата работа

Избрахме тази задача след преглед на скорошните предизвикателства в kaggle. На пръв поглед тя ни грабна вниманието и я избрахме единодушно. Началните ни идеи се оказаха неверни и след обмисляне, тръгнахме в напълно различна посока.

Източник на задачата: https://www.kaggle.com/c/poker-rule-induction. Даден е train.csv файл, в който има хиляди ръце на покер. На всяка от тях е дадена стойност/клас от 0 до 9. Има и друг файл – test.csv, в който са дадени милион други ръце без стойност и именно тези стойности трябва да се познаят на база на предишните ръце.

2. Кратък обзор

Задачата е от класификационен тип и има множество подобни на нея. Основната разлика, е че тук самите атрибути и техният вектор не са от значение, а по-скоро отношението между тях. Има множество предложения за решения на различни езици на конкретно тази задача, но не нашето не е повлияно от нито едно от тях.

Подходящ метод за решаване е формирането на Decision tree след детайлно анализиране и запознаване с входните данни. Както е предложено в условието на задачата (а както и ние сме подходили) е добре да се използва RAGA – техника, при която правилата са по-генерални. Те се измерват с 2 показателя: accuracy и coverage – в каква част от случаите за даден клас правилото е изпълнено и в колко други класове правилото отново е изпълнено. Най-ценни са правилата, които са верни за максимален процент от примерите за даден клас и минимален процент за примерите от всеки друг клас. Тоест тези, които могат да направят разликата.

3. Нашето решение

Нашето решение е следното: считаме за дадено, че при игрите на карти, подредбата няма значение. „Подсказваме“ на програмата какво да търси: поредни карти, карти от една и съща боя, карти с една и съща стойност и т.н. Дефинираме такива правила за всяка един от класовете ръце и анализираме резултатите. След като става ясно, че правилата не са достатъчно изчерпателни и не дефинират точно даден клас, навлизаме в детайли за всяко едно от тях. Например:

- колко са поредните карти; от какви бои са; коя карта е началната
- колко са еднаквите по боя карти; от каква боя са; с какви стойности са; тези стойности поредни ли са
- с каква стойност са еднаквите по стойност карти; колко са на брой;
- колко пъти в една ръка се среща всяко едно правило; в колко процента от ръцете от даден клас това правило се повтаря

Уверихме се, че за да стигнем до добри заключителни резултати, трябва да навлезем в пълни детайли и чак тогава можем да разграничим една ръка от друга. След много опити, пренаписвания и поправки на реализацията стигнахме до редица заключителни правила за всеки един клас ръце.

4. Програмна реализация

Реализирахме програмата на .NET конзолно приложение. След това направихме и уеб приложение, което чете и визуализира получените резултати. То е с демонстративна цел.

Реализацията протича на няколко етапа:

- Прочитане на входните данни и запазване на информацията от тях в Dictionary – за всеки клас има списък от ръце
- Извличане на поредица от правила за всяка една ръка
    - Подобряване и детайлизиране на правилата
    - Определяне дали всяко едно правило е заключително и дали се отнася за всички ръце от този клас
- Определяне на финални правила и дефиниране на метод, който спрямо тях изчислява класа на произволна ръка
- Изпробване на метода върху тестовите данни
- Записване на резултатите в нов файл с цел използване от уеб приложението

5. Заключение

В крайна сметка постигнахме целта си. Програмата прочита тренировъчните данни и определя предварително дефинирани типове правила към всеки един клас. Те са достатъчни, за да се определи всяка следваща ръка в кой клас ще попадне. Тествахме реализацията върху тестовите данни и подготвихме уеб интерфейс, който да демонстрира постигнатото.

Проектът може да търпи множество бъдещи подобрения и разширения. Може да се смени подхода и да не се търсят строго дефинирани правила, за да стане по-универсална програма. Може да се изпробва и върху други игри с карти и да се използва в приложения за игра на карти срещу машина.

6. Литература

http://www.wseas.us/e-library/conferences/crete2002/papers/444-494.pdf