

映画興行収入の予測とスタッフ及びキャスト間の相関ルール発見

名列番号:032 菅原勝彦

1.はじめに

統計解析ソフト R を用いて、普段あまり利用しない統計解析について身の回りのデータを使って実際に処理を行い、理解を深める。映画興行収入の予測とスタッフ及びキャスト(以下 SC)間の相関ルールを発見し検証する。

2.研究に用いたデータ

本実験では“社団法人日本映画製作者連盟”と“映連 DB”から抽出した映画作品とその興行収入、及び SC の情報を解析用に編集して用いた。

<http://www.eiren.org/index.html>

予測と相関ルールに使用するデータは互いに異なる。予測に用いるデータは、行の要素に映画作品、列の要素に SC と興行収入の値を持った 325 行 3003 列の表となる(サイト内の SC 掲載順に、仮に先頭が重要度の高いものと仮定して、1~0.5 の重み付けがされている)。相関ルールに用いるデータは、映連DBに存在する全映画作品毎に SC がいるか、いないかを 0 と 1 で表した 752 行 15,000 列の表となる。

3.実験内容

・予測・・・与えられたデータ群から、近似式を導出し、近似式に値を代入することで興行収入を予測する。R では、パッケージ e1071 内の関数 SVM による線形回帰を用いて予測を行う。

・相関ルール・・・与えられたデータから、頻出する組み合わせについて、支持度、確信度のパラメータによりルールの評価を行う。R では、パッケージ arules 内の関数 APRIORI を、最小支持度=0.05/最小確信度=0.01 で用いる。

4.結果

SC のデータだけでもある程度興行収入の予測は可能であった。結果の一部を図 1 に示す。この結果から分かることは、興行収入の平均 20.9 億円に対して極端に高い作品は、予測値が大幅に外れるということである。予測には対象の作品以外のデータを用いるため、このような誤差が生じてしまうと考える。tf-idf で重みを大きくすることで、予測精度が向上していることが図 2 より確認できる。だが、踊る大捜査線シリーズはお互いに SC がほとんど同じである為、一方の予測を行っているときはもう一方の予測値に近いものを出力するという問題点が確認できた。

発見されたルール(図 3)を見ると、全 48 作品ある”男は辛いよ”シリーズ関連のスタッフが多く発見され、全体を通して確信度が 100%に近いものが多いことが分かった。これはデータ内に、 $A \Rightarrow B$ というセットが常に成り立っているということであり、シリーズモノでは SC の変化があまり無く、それ以外の作品では、その一作品のみでの関係であると言うことができる。列に対する行の要素数を増やすことが出来れば、発見される相関ルールのバリエーションも増えるだろう。

5.おわりに

統計解析に触れてみて、解析に取り扱い易いデータを用いることが一番重要だと感じた。今回のデータは、要素のほとんどの重みが 0 だったことと、列に対する行の要素数が少なすぎた点において、あまり解析には適していなかったと思う。今回は過去のデータを用いた予測であったが、過去のデータが無い状態での予測には何が関与するのかを考えるのも興味深いと思う。今回の自主課題研究を通してデータがあれば、身近な事柄についての予測や、相関ルール発見に役立てたいと思った。

タイトル	興行収入	予測値	誤差	誤差の絶対値
三等高校生&&ジェ	11	11.00689	0.00689	0.006893
北の零年	27	27.02001	0.02001	0.020014
ゴジラVSメカゴジラ	18.7	18.6497	-0.0503	0.050303

もののけ姫	113	30.29063	-82.7094	82.70937
踊る大捜査線 TH	173.5	39.78278	-133.717	133.7172
ハウルの動く城	196	26.22329	-169.777	169.7767
千と千尋の神隠し	304	27.86554	-276.134	276.1345

図 1.予測結果

タイトル	興行収入	予測値	誤差	誤差の絶対値
NIN NIN 忍者ハッ	19.3	19.32035	0.020346	0.020346
東京タワー ボクと	18.8	18.77249	-0.02751	0.027511
名探偵コナン 瞳の	25	25.04634	0.046343	0.046343

踊る大捜査線 TH	50	159.4351	109.435	109.4351
踊る大捜査線 TH	173.5	40.38155	-133.12	133.1185
紅の豚	28	172.8623	144.862	144.8623
千と千尋の神隠し	304	88.12253	-215.88	215.8775

図 2.tf-idf による予測結果

1 {監督_山田洋次} => {脚本_山田洋次}	0.07928389	1.000000
2 {編集_石井巖} => {脚本_山田洋次}	0.07544757	0.8550725
3 {監督_山田洋次} => {編集_石井巖}	0.07289003	0.9193548
4 {編集_石井巖} => {監督_山田洋次}	0.07289003	0.8260870
5 {監督_山田洋次, 編集_石井巖} => {脚本_山田洋次}	0.07289003	1.000000

図 3.相関ルール