6R-1

映画の興行収入を予測する手法の検討

CHANGVISOMMID LINDA† 青野 雅樹 †
† 豊橋技術科学大学情報・知能工学系

1 はじめに

映画産業において顧客からのフィードバックを知ることが重要である．映画の興行収入は，その映画を見た顧客からのフィードバックの一つの重要な指標でもある．映画が公開される前に，ある程度映画の興行収入を予測できると，映画スタジオや配給会社などがその映画を製作するあるいは買うかどうかの決断の支援につながるであろう．

本研究では，映画の特徴を発見できる素性を提案し，映画の興行収入を予測する手法の検討を行う．

2 関連研究

映画の興行収入を予測する研究では，予測問題を分類問題に帰着し，ニューラルネットワークを用いた興行収入の分類が行われた．

Sharda ら [1] は，834 本の映画を取得し，映画が公開させる前を前提とし，26 素性を抽出した．また，映画の興行収入を 9 カテゴリに分けて，映画はどのくらい儲かるかを知るために分類実験を行った．彼らは 10-交差検証でシステムを評価した．その結果，該当カテゴリに正確に分類したのは約 36.9%で，隣り合うカテゴリを含めた場合の正解率は約 75.2%であった．

3 データ

本研究で用いた映画のデータは「IMDb」という映画のサイトから取得し，映画の興行収入の情報は「Box office mojo」というサイトから取得した．その映画のデータは，2006 年から 2010 年までの期間のものである．今回の実験では，1893 本の映画を用いた．

それらの映画を 9 カテゴリ (Hit1「大失敗」∼Hit9「大ヒット」) に分けて，映画の数がほぼ同数となるように興行収入の基準を選択した．ただし，今回の実験で用いた映画の興行収入は，ビデオレンタル，関連グッズやサウンドトラックの販売などの収入を含まないとする．実験で用いた映画のカテゴリを表 1 に示す．

表 1: 映画のカテゴリ

カテゴリ	金額 $x(\times 10^6$US\$)	映画の数
Hit1	$x \leq 0.18$	203
Hit2	$0.18 < x \leq 0.7$	214
Hit3	$0.7 < x \leq 2$	208
Hit4	$2 < x \leq 6$	211
Hit5	$6 < x \leq 14$	216
Hit6	$14 < x \leq 31$	211
Hit7	$31 < x \leq 68$	219
Hit8	$68 < x \leq 150$	205
Hit9	$x > 150$	206

表 2: 提案素性

番号	素性	種類
1	ジャンルの値	実数値
2	出演者の値	実数値
3	監督の値	実数値
4	上映時間	整数値
5	公開国の数	整数値
6	小説ベース	Boolean
7	続編	Boolean

4 提案素性

提案素性を表 2 に示す. 今回提案する素性は，Boolean 素性と実数・整数値素性の 2 種類がある．Boolean 素性は，映画の情報の中に指定したものを含んでいるかどうかを 2 値で表す素性である．実数・整数値素性に関しては，それぞれの映画に含まれる情報によって値が異なる．

ここで，定義したジャンルの値は，それぞれのジャンルに対して映画が公開された年から遡り過去 5 年間の興行収入の平均値から求めたものである．ジャンルが複数ある場合，それぞれのジャンルの値の和を求めて，一つのジャンルの値とする．

出演者の値および監督の値は，ジャンルの値を求めると同様に，過去 5 年間のの興行収入の平均値から求め，出演者または監督の数で割り算して，それぞれの映画の出演者の値・監督の値とする．

A study on predicting box-office success of movies
CHANGVISOMMID LINDA†, Masaki AONO†
†Dept. of Computer Science and Engineering, Toyohashi University of Technology
441-8580, Toyohashi, Japan

表 3: 10-交差検証による実験結果

		実測値								
		Hit1	Hit2	Hit3	Hit4	Hit5	Hit6	hit7	Hit8	Hit9
予測値	Hit1	87	57	34	18	22	13	7	6	1
	Hit2	51	68	54	44	36	15	13	4	0
	Hit3	11	15	19	20	13	8	4	1	0
	Hit4	10	23	38	52	30	16	11	2	1
	Hit5	19	32	38	38	55	40	15	5	0
	Hit6	8	12	17	22	35	58	34	16	2
	Hit7	8	2	6	11	21	50	81	59	17
	Hit8	3	2	2	3	3	7	36	54	45
	Hit9	6	3	0	3	1	4	18	58	140

5 評価実験

5.1 実験概要

提案素性の有効性を確認するために、評価実験を行う．実験データとして 2006 年から 2010 年の間に公開された 1893 本の映画を用いた．そのデータをランダムに取り出し，興行収入の分類を行う．本実験では，SVM と Random Forest という機械学習を用いて実験する．

実験の手順として、まず実験データを訓練データセットと，テストデータセットを作成する．ここで，10-交差検証で，データの数をだいたい等しくなるように 10 等分のグループに分割する．その後、SVM で 9 グループを訓練し、訓練したモデルを 1 グループのテストデータ用に当てる．この作業を繰り返し，平均正解率を求めて，システムを評価する．

5.2 実験結果

SVM 分類器で訓練し、10-交差検証によりテスト用のデータに当てた結果を表 3 に示す．ここで，該当するカテゴリのみに分類した正解率と隣り合うカテゴリを含めた場合の正解率について求めた．実験結果から，該当するカテゴリのみに分類した正解率は，約 32.4%であることがわかった．一方，隣り合うカテゴリを含めた場合の正解率は，約 67.3%であった．また，各カテゴリ間について正解率を見ると，最も正解率が高かったのは，Hit9「大ヒット」に分類した結果 (該当するカテゴリのみに分類した正解率：67.9%) であることがわかった．

さらに，提案素性の中でどの素性が最も有効であるかを調べるために，Random Forest を利用した．素性の有効性を調べた結果を図 1 に示す．一番効き目が良かったのは公開国の数の素性で，その次が出演者の値の素性であることが分かった．

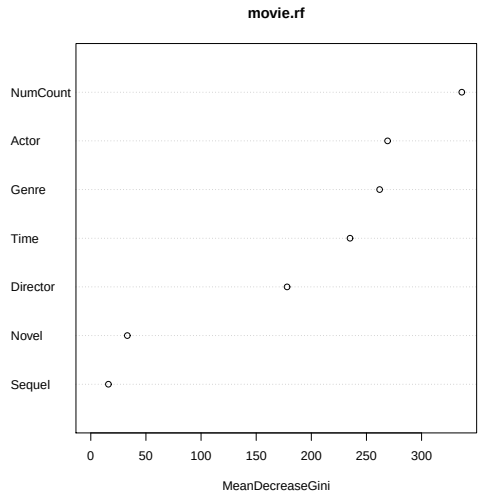

図 1: Random Forest から算出された重要素性

6 おわりに

本研究では，映画の特徴を発見する素性を提案した．今回の実験結果で公開国の数や出演者，ジャンルは，映画の興行収入に影響を与えることがわかった．しかし，映画の興行収入を 9 カテゴリに分類するには提案素性のみでは不十分であると考えられる．

今後の課題として，より精度の高い素性を検討し，映画がしばらく公開された後，ユーザからレビューを用い，映画の興行収入に影響を与えるを解析していきたい．

参考文献

[1] Ramesh Sharda, Dursun Delen: “Predicting box-office success of motion pictures with neural networks”, *Expert Systems with Applications*, Vol.30, pp. 243-254, (2006).

[2] Li Zhang, Jianhua Luo, Suying Yang: “Forecasting box office revenue of movies with BP neural network”, *Expert Systems with Applications*, Vol.36, pp. 6580-6587, (2009).