COMP90049 Project 2 Report

Abstract

This project is aimed to build a movie genre classification model by using one of several supervised machine learning algorisms. The model will analyze movies' features such as titles, years and audios, etc. Then it will provide a predefined movie genre based on these features. Each model will be accessed and criticized by their performance over the test dataset.

1 Introduction

Nowadays machine learning is become more and more popular. Big companies such as Amazon and Google have used machine learning models to predict their customer preference. As a result, these companies are able to gain more profit with their accurate products' advertising. In media industry, data and information is valuable as well. Movie companies need market information to make decision on developing their film products such as their target population and popular movie genres, in order to attain more income. The stream media company such as Netflix uses algorisms to anticipate their users' favourite movie genres and predict what they probably desire to watch in the next.[1] So, they will have no need to spend time to browse the whole website and look for a movie they prefer. This kind of services is able to improve users experience by saving their time and effort. As a result, old customer loss is prevented, and the user groups' amount is maintained at a certain level, in addition new customers possibly will be attracted to join in.

In this project a movie genres' classification model is built by utilizing several machine learning methods. Movies' features such as titles, audio data, video data, viewer comments are collected and used by the model to perform predictions. The prediction's result will tell which movie genre the movie is classified. After that, the accuracy of each model will be compared and then the model with highest accuracy will be chosen.

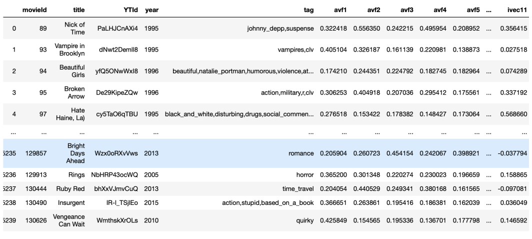
2 Literature Review

Stephan Dreiseitl [3] pointed that there are many implementations of movie classification software already exist. Their performance is primarily decided by three aspects, how good is the data sets' quality, how suitable are the models chosen, and the evaluation methods used in the result. Most of nowadays study focus on two problems, is the model able to categorize objects with different classes and can the model predict the object with unfamiliar features. In 2019 Haili Zhang's research says that there are some limitations of content-based classifier. One of them is the fact that the algorism relies on appropriate data set, which means it can not provide accurate result if the data set has no sufficient information. In her research SVM and logistic regression is used in the model. The performance of the model is quite successful, however the training data set chosen has some redundant and the accuracy is negatively affected.

3 Methodology

The program language used in this project is python 3.

2.1 Data pre-processing



	movieId	title	YTId	year	tag	avf1	avf2	avf3	avf4	avf5	...	ivec11
0	89	Nick of Time	PaLHJCnAXi4	1995	johnny_depp,suspense	0.322418	0.556350	0.242215	0.495954	0.208952	...	0.356415
1	93	Vampire in Brooklyn	dNwt2DemlI8	1995	vampires,clv	0.405104	0.326187	0.161139	0.220981	0.138873	...	0.027518
2	94	Beautiful Girls	yfQ5ONwWxI8	1996	beautiful,natalie_portman,humorous,violence,at...	0.174210	0.244351	0.224792	0.182745	0.182964	...	0.074289
3	95	Broken Arrow	De29KipeZQw	1996	action,military,r,clv	0.306253	0.404918	0.207036	0.295412	0.175561	...	0.337192
4	97	Hate Haine, La)	cy5TaO6qTBU	1995	black_and_white,disturbing,drugs,social_commen...	0.276518	0.153422	0.178382	0.148427	0.173064	...	0.568660
...	...	...	...	...	...	...	...	...	...	...	...	...
5235	129857	Bright Days Ahead	Wzx0oRXvVws	2013	romance	0.205904	0.260723	0.454154	0.242067	0.398921	...	-0.037794
5236	129913	Rings	NbHRP43ocWQ	2005	horror	0.365200	0.301348	0.220274	0.230023	0.196659	...	0.158865
5237	130444	Ruby Red	bhXxVJmvCuQ	2013	time_travel	0.204054	0.440529	0.249341	0.380168	0.161565	...	-0.097081
5238	130490	Insurgent	IR-I_TSjIEo	2015	action,stupid,based_on_a_book	0.366651	0.263861	0.195416	0.186381	0.162039	...	0.036049
5239	130626	Vengeance Can Wait	WmthskXrOLs	2010	quirky	0.425849	0.154565	0.195336	0.136701	0.177798	...	0.146592

Figure 1- The Original train data

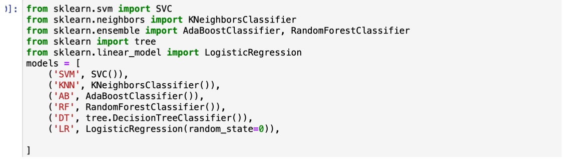
Figure 1 shows the original data of this project. The raw data provided is not in a standard format, some data was missing, and some columns of data is in different shape, for example some of the year information is shown as integers as '1996', some of it are shown as '2000)'. To fix this problem, Microsoft Excel is used, and the different form data is changed to a standard format manually.

Some features are not helpful in performing analysis. Movie ID and YTID are unique features, which means each object has different value. With unique features, the machine learning model's accuracy will be influenced. As a result, these features were moved in the data pre-processing step.

Before import the feature data into the learning model, some of the features are not acceptable by the model. For example, the Naïve Bayes model in sklearn does not accept string data, so LabelEncoder method, CountVectorizer method was applied to features with string format.

2.2 Build Model

Several classifiers are used in this project, such as decision trees classifier, Gaussian Naïve Bayes classifier and logistic regression classifier. In figure 2 the code of the model is provided.



```
]: from sklearn.svm import SVC
   from sklearn.neighbors import KNeighborsClassifier
   from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier
   from sklearn import tree
   from sklearn.linear_model import LogisticRegression
   models = [
       ('SVM', SVC()),
       ('KNN', KNeighborsClassifier()),
       ('AB', AdaBoostClassifier()),
       ('RF', RandomForestClassifier()),
       ('DT', tree.DecisionTreeClassifier()),
       ('LR', LogisticRegression(random_state=0)),

   ]
```

Figure 2- Code of several classifiers' model

4 Algorisms

3.1 Naïve Bayes Algorism

The first algorism chosen is Gaussian Naïve Bayes algorism.

Naïve Bayes model is naïve is because it is established based on the assumption that each features of the object are not related to each other [2]. The probability of event Y based on event X can be calculated by the formula in figure 3.

$$P(x,y) = P(y|x)P(x) = P(x|y)P(y)$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Figure 3- Bayes Rule formula

Because of each feature of the objects feature is not related, the probability of the movie genre based on given features can be written as the formula in figure 4.

$$P(x,y) = \prod_{i=1}^{N} P(y^i) \prod_{m=1}^{M} P(x_m^i|y^i)$$

Figure 4- Naïve Bayes model's formula

After calculating each genres probability of each genres depend on the given features of the single movie, the model will classify the movie into the genre which has the highest probability.

The advantage of Naïve Bayes algorism is the fact that it can be easily built and estimated. In addition, it has the ability to be scaled to many different dimensions.

3.2 K-Nearest Neighbors Algorism

In K-nearest neighbors algorism, the data is directly used to classify the movie's genre [3]. The object is classified according to the majority class of the K-Nearest training instances. As a result, in machine learning model's construction procedure, there is no parameter to be considered except the 'K'. 'K' means how many neighbors are included in estimating the movies' classes membership, which means the flexibility of the model will depend on the value of k.

The main superiority of this algorism over others is that evidence can be provided by the neighbors to explain the final result. However, it has a disadvantage that the calculation of the distance between data items is not effortless and sometimes it is even impossible.

3.3 Decision Trees

This algorism split the data into a tree like structure, each node of the tree implies an attribute, the most common class of the subset will be used to accredit each leaf of the tree. It will iterate the splitting procedure until it finds the most suitable tree to perform the prediction. The structure of decision trees is displayed in figure 5.

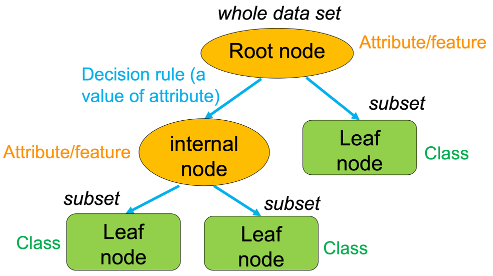


Figure 5- Decision tree's structure

The advantages that decision trees have is that it is quite easy to understand and apply, and an exhaustive analysis of the aftereffect along each branch is created by it. However, it has the drawback that due to the greedy algorism used during tree construction, continuous variables will be completely discretized [4], and this will contribute to the information loss of the whole feature dataset. Furthermore, it will reduce the accuracy of the model.

3.4 Logistic Regression

In logistic regression model, the probability of the result is decided by the formula $P(Y|X) = f(x, a)$. After that, a decision boundary is defined. So, if the value of the function is lower than it, the result will be zero, otherwise the result will be one. The function graph is shown in the figure 6. The parameter *a* is obtained by minimizing the negative conditioning log likelihood.

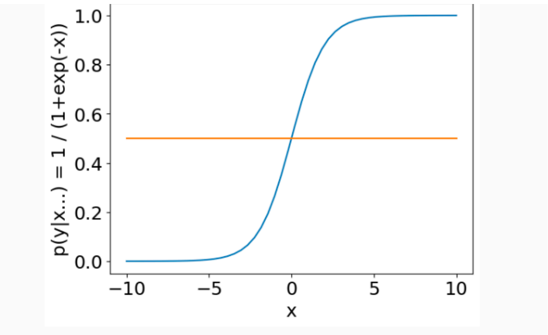


Figure 6- Function graph of P(y|x)

The advantages of logistic regression is the fact that it performs well with frequency-based features and in most of the time it performs better than Naïve Bayes model. In addition, it does not obligate conditional assumption on the features. The main disadvantage of this algorism is that usually the model needs considerable feature data to perform properly. And it has the limitation that it is able to review linear connection of features' data

5 Evaluation

After applying each model to the train dataset, the accuracy values of the models are obtained. The result is shown in the table below.

model	accuracy
Naïve Bayes	0.157
SVM	0.207
K-Nearest N	0.160
Decision Tree	0.221
Logistic Regression	0.408

Table 1- accuracy of each models

As it is observed from the table Naïve Bayes model has the worst performance. The main reason leads to this result is that all the features of a movie actually have relationships which means each feature are connected and this is in contrast of the naïve assumption.

K-nearest neighbours model has a poor performance as well. It is because this model can only have considerable performance in small dataset and low dimensions.

The accuracy of decision tree's model is not quite high. During the model construction the continuous data which in this case is video and audio data are discretised. As a result, the feature data suffered information loss, and this leads to the drop of the accuracy.

Logistic regression model has the best performance overall. It shows that this model can work well on large amount of feature data. The accuracy of it can be increased if a more proper threshold is applied.

6 Conclusions

In this project several machine learning classifiers are used, the results of these models are compared, and finally a model with the best performance is obtained which is logistic regression model. In the report the algorisms used are explained and evaluated. The advantages and disadvantages of these algorisms are described and clarified.

However there are still many factors haven't be considered in this project, which means the models are still possible to be improved to gain higher accuracy and perform better.

References

Note that the recommended citation style for this report is defined as Harvard style, but you may use other (formal) citation styles if you prefer.

[1] Haili Zhang, 2019. Movie Genre Preference Prediction Using Machine Learning For Customer-Based Information.

[2] Dr.K, U. and Dr.M, K., 2020. Performance Analysis of Naïve Bayes Correlation Models in Machine Learning. International Journal of Psychosocial Rehabilitation, 24(04), pp.1153-1157.

[3] Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5-6), pp.352-359.

[4] AKSU, G. and KECEOGLU, C., 2019. Comparison of Results Obtained from Logistic Regression, CHAID Analysis and Decision Tree Methods. Eurasian Journal of Educational Research, 19(84), pp.1-20.

[5] Deldjoo, Y., Elahi, M., Quadrana, M. and Cremonesi, P., 2018. Using visual features based on MPEG-7 and deep learning for movie recommendation. International Journal of Multimedia Information Retrieval, 7(4), pp.207-219.

[6] Harper, F. and Konstan, J., 2016. The MovieLens Datasets. ACM Transactions on Interactive Intelligent Systems, 5(4), pp.1-19.