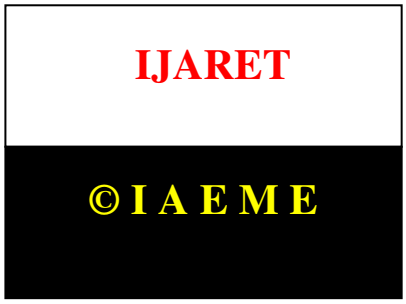
APPLYING SUPERVISED AND UN-SUPERVISED LEARNING APPROACHES FOR MOVIE RECOMMENDER SYSTEM

Jai Prakash Verma[1], Rasendu Mishra[2], Rushabh Shah[3], DevendraVashi[4]

Volume 6, Issue 6, June (2015), Pp. 18-22
Article ID: 20120150606004
International Journal of Advanced Research in Engineering and Technology (IJARET)
© IAEME: www.iaeme.com/ ijaret.asp
ISSN 0976 - 6480 (Print)
ISSN 0976 - 6499 (Online)



[1,2,3,4]Assistant Professor, CSE Department, Institute of Technology,
Nirma University, Ahmedabad

ABSTRACT

In this research paper we are trying to compare supervised and un-supervised machine learning approaches for comparing to identify the necessity of these techniques for developing a recommender system for movies. Also we are trying to adjust the training and testing samples to get the best accuracy for the recommender system.

Keywords: Machine Learning, Classification, Clustering, Recommender System

I. INTRODUCTION

In today's era, consumers are more wise and they need a system which can guide them for utilizing any market product based on the preferences given by the others who had used them. Hence, we can say that purchases are based on reviews given by others. For this we need an automated Machine Learning based Recommender system which can recommend and guide users in many different ways as possible to make them reach to suitable and optimal judgement for utilizing any marked based product or services[7][8]. The most commonly used Machine Learning techniques are classification and clustering whereby we would classify or cluster the recommendations made by others and use them for future decision making. In developing country like India where E-Commerce is becoming more and more utilized for buying and selling goods and services, this recommender system may be of great usage for the consumers and sellers. A survey shows that e-commerce activity is going to be of $8.5 billion by 2016 and there the profit and losses may be dependent on a Machine Learning based recommender systems [4].

II. WHY RECOMMENDER SYSTEM

A recommender systems would be an automated software system which would be trained to make decisions intelligently for new and future inputs. The training is given to the intelligent software system using prior events and their outputs which exists in the world history as facts now. Once the system is trained, the users can give inputs for the future problems and the system would, after some data mining techniques operations, would give you the best suited results. In this research paper, we had tried to compare the results of Movielens database system obtained after applying data mining techniques like Clustering and Classification Movie Ratings and tried to create a

recommender system as to which movies are rated high and low by the users in the scale of 1 to 5 where 1 means poor and 5 means excellent. Once this recommender system is fully established then it can intelligently recommend the consumers whether a movie is viewable or not [7] [8].

III. BASICS OF CLASSIFICATION

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of object whose class label is unknown. The derived model is based on the analysis of set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Hence the output of the classification depends upon the types of classifier used[2][3].In our case,we have used probabilistic approach to find the classification of likes and dislikes of movies, which is represented in terms of probabilities of success or failure.

IV. BASICS OF CLUSTERING

There exist various clustering methods varying due to the method of finding similarity between clusters as some use cosine methods or Euclidean distance measures or Manhattan distance measure resulting in to clusters of variety of shapes. But basically we can cluster in to two general ways:

- Partitioning Approach of clustering: Here we think that all item sets which are tobe clustered are in one big cluster initially then we start finding similarity between the data-items and form sub-clusters and further repeat the process of finding similarity with in each of those sub-clusters until we get a cluster with most similar item-sets. Here we use Top-Down approach for clustering and called as divisive approach of clustering in hierarchy [1] [2].
- Agglomerative Approach of Clustering: Here we assume that all item sets areindividual. We try to find similarity between item-sets and try to merge only those item sets which are similar and thereby form a root. Again these roots are used to find similarity amongst them and other un-clustered data-items (also called clusters) and again superior roots are formed. This approach is called Agglomeration approach.[1][2]

V. PROBLEM UNDER STUDY

i. Apply clustering based techniques to develop the recommender system.
ii. Identify how much changes percentage selection of training dataset would result in best test results using various classification algorithms so as to develop best classification based recommender system.

VI. DETAILS ABOUT EXPERIMENT AND SCOPE OF THE EXPERIMENT

- We are using internationally accepted data about MovieLens data set.
- For supervised learning the methods used are: NaiveBaise Classification, DT-J48, Rule based decision tree classification and Decision Stump Classification.
- For un-supervised learning the methods used are: K-Means Clustering
- For programming and experimentation, the tool used is R-package and Weka [5][6].

- The experimentation is performed on Intel® Core™ i3-3110M CPU at 2.40GHz, 64 bit and 4GB RAM.

VII. EXPERIMENTAL RESULTS

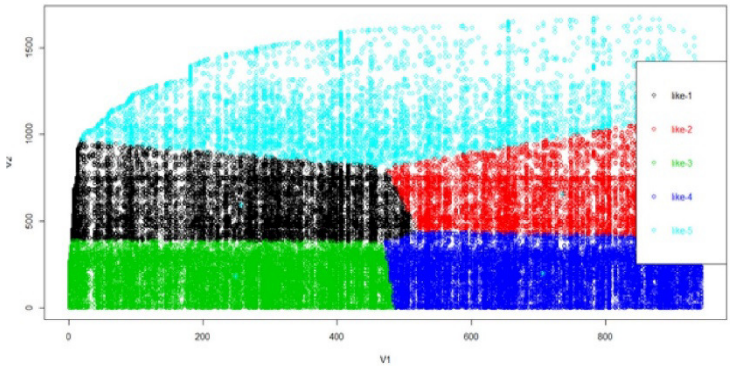


Fig 1: Clustering using K-Means based on feature "MovieLikes".

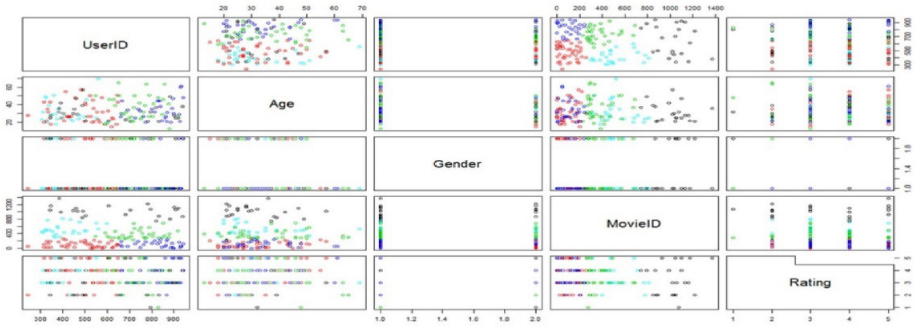


Fig 2: Clustering using K-Means based on different features

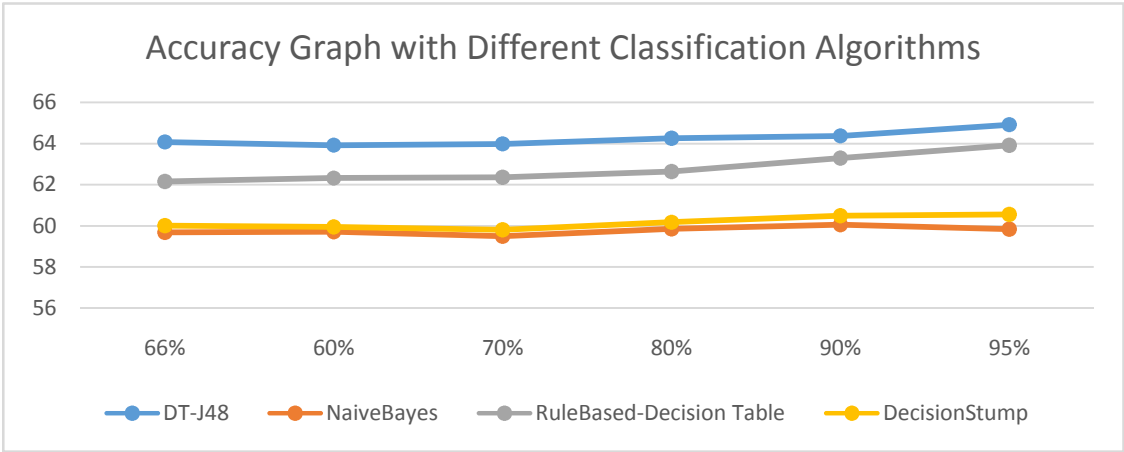


Fig 3: Classification Results

Table 1: Percentage accuracy table for different classification algorithms for movies recommendation model						
Algorithm	**Percentage split of dataset in Training Dataset and Test Dataset**					
	66%	**60%**	**70%**	**80%**	**90%**	**95%**
DT-J48	64.0824	63.925	63.9867	64.26	64.38	64.92
NaiveBayes	59.6824	59.72	59.5	59.86	60.06	59.84
RuleBased-Decision Table	62.1588	62.325	62.36	62.64	63.3	63.92
DecisionStump	60.0176	59.955	59.8133	60.18	60.5	60.56

Fig 4: Comparative Study table for various Classification Algorithms after varying training data percentage and test result accuracy.

VIII. INFERENCES FROM THE STUDY

- Fig 1 represents the results obtained after application of K-Means clustering on Movielens data. The different colours shows the viewablility of a movie based on recommendations made by rating the movies by various users. Hence, based on this graph a person can take decision whether to view the movie or not.
- Fig 2 also represents the results obtained after applying K-Means clustering algorithm on Movielens data. This figure is showing clusters on movie rating based on different features of users like Age, Gender etc.
- Fig 3 represents the results obtained after application of various classification techniques on Movielens data. The graphical representation of fig 4 are summarized in fig 3 table which clearly denoted that if we keep training data up to 80% the test results are more accurate.

IX. CONCLUSION

Form this study we can conclude that test results for Movielens Recommender System would give good results if the nearly 80% data is taken for training the system. Also we can use Clustering Approach for identifying the viewablility of a movie based on MovieRatings. Hence, such Recommender systems, if built with intelligence using Machine Learning, may bring the future decisions more and more closer to human thinking and may even sometimes go beyond human thinking limits.

REFERENCES

1. Mishra R., Modi N., "A Novel Approach to cluster new data-items in previously clustered data-items using Agglomerative Clustering with Single Link.
2. Jiawei Han and MichelineKamber. Data Mining: concept and Techniques. ElsevierPublication.
3. Jai Prakash Verma, SapanMankad, "Smart Inbox: A comparison based approach to classify the incoming mails", International Journal of Artificial Intelligence and Knowledge Discovery Vol.1, Issue 1, Jan, 2011
4. Indian E-Commerce Stats: Online Shoppers &Avg Order Values To Double In Next 2 Years!, Web Link" http://trak.in/tags/business/2014/04/04/indian-e-commerce-growth-stats/f" on dated 20-06-2015
5. "Weka 3: Data Mining Software in Java", Web Link: http://www.cs.waikato.ac.nz/ml/weka/ on dated 20-06-2015

6. W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.2.0 (2015-04-16)
7. Jai Prakash Verma, Bankim Patel, Atul Patel,"Big Data Analysis: Recommendation System with Hadoop Framework", 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 978-1-4799-6023-1/15, 2015 IEEE DOI 10.1109/CICT.2015.86
8. Sandra Garcia Esparza, Michael P. O'Mahony, Barry Smyth, "Contents lists available at SciVerseScienceDirect Knowledge-Based Systems", Knowledge-Based Systems 29 (2012) 3–11.
9. Priyank Thakkar, Samir Kariya and K Kotecha, "Web Page Clustering Using Cemetery Organization Behavior of Ants" International Journal of Advanced Research in Engineering & Technology (IJARET), Volume 5, Issue 1, 2014, pp. 7 - 17, ISSN Print: 0976-6480, ISSN Online: 0976-6499.
10. Neeti Arora and Dr.Mahesh Motwani, "A Distance Based Clustering Algorithm" International journal of Computer Engineering & Technology (IJCET), Volume 5, Issue 5, 2014, pp. 109 - 119, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.