International Journal of Latest Trends in Engineering and Technology
Vol.(11)Issue(1), pp.059-063
DOI: http://dx.doi.org/10.21172/1.111.11
e-ISSN:2278-621X

TOP N RECOMMENDATION SYSTEM USING SELECTABLE RANDOM AND HYBRID MACHINE LEARNING TECHNIQUES

Abhijit D. Awari[1], Dr. Mahesh R. Dube[2]

Abstract- Recommendation systems(RS) attempt to recommend the most suitable item to users by using different predictive algorithms. Recommendation systems are used to perform three main tasks. These tasks include rating prediction in which RS aims to fill the missing entries in User-Item Rating Matrix, Top N recommendation in which system generates a ranked list of N items to users, Classification in which items are classified into correct categories. In this paper, we have proposed a generic model for a Top N Recommendation system using Selectable Random an Hybrid Machine Learning Techniques. We have implemented a data pre-processing module and a classification module using different machine learning techniques. The proposed system generates the Top N recommendations. We have used the Wine data set which is available on Internet. The proposed system can be tweaked to generate the recommendations according to the data set. Experimental results show that how our recommender system works on an example dataset.The system is evaluated using different evaluation metrics.

Keywords –Recommendation System(RS), Machine Learning, Classification, Data Preprocessing, KNN, PCA.

1. INTRODUCTION

In recent years, the recommendation systems are useful to predict something which is most appropriate and relevant for an individual user. The recommendation system is a part of the information filtering system which removes unwanted information from an information stream prior to presentation to a human user. Generally, the recommendation systems use content based filtering or collaborative filtering or combination of both(hybrid).

The rest of the paper is organized as follows. A formal model and block diagram of the proposed system are explained in section II. Experimental results are presented in section III. Concluding remarks are given in section IV.

2. PROPOSED SYSTEM

2.1 Formal Model –

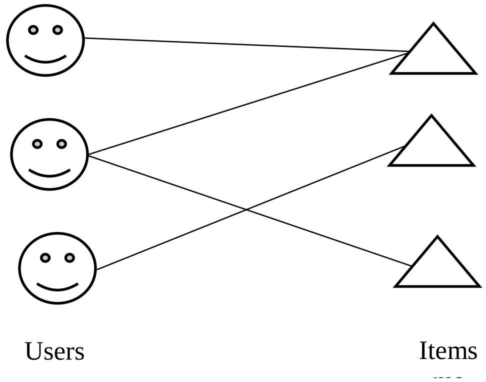
We can represent a Recommendation System using a bipartite graph. A bipartite graph is a graph with its vertices partitioned into two disjoint sets.A formal model for a recommendation system is as follows,

$Y =\{(u, i), (u,u), (i , i) : \text{ where } u \in U, i \in I\}$

The two disjoint sets are as follows,

$U = \{U1, U2,..., Um\}$ = A set of ‘m’ users

$I = \{I1,I2,\dots, In\}$ = A set of ‘n’ items



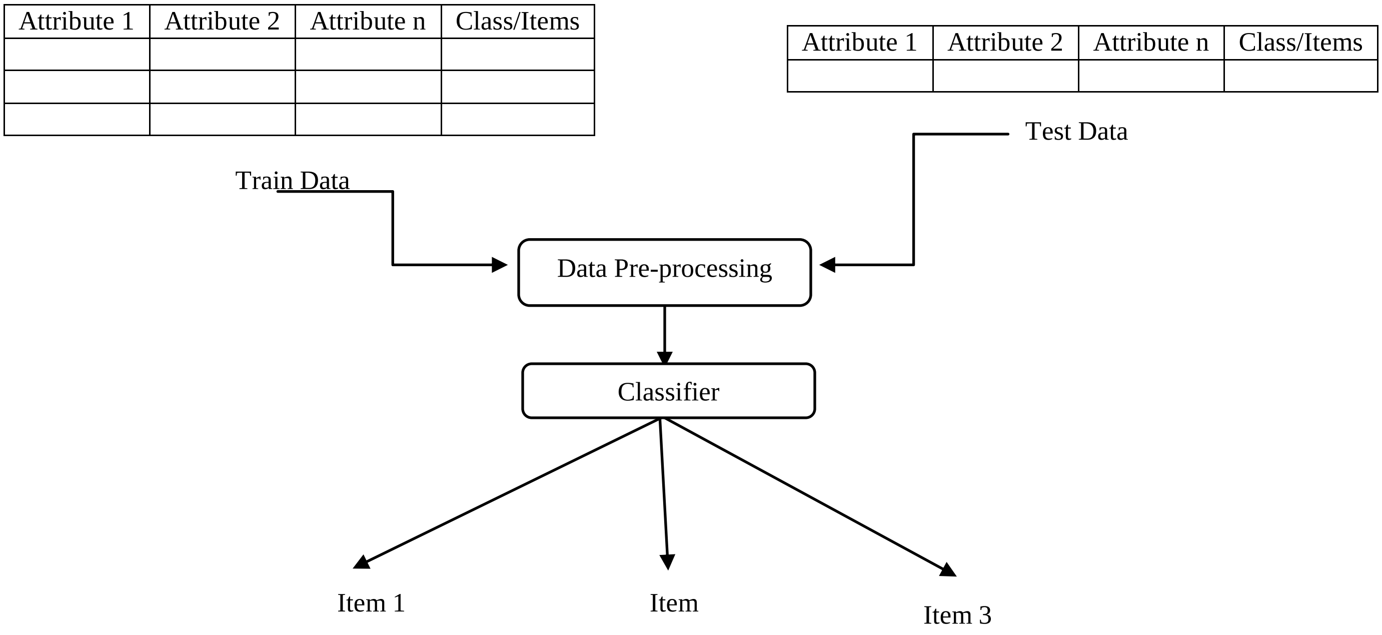
A formal model for a recommendation system

[1] Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, India

[2] Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, India

2.2. Block Diagram –

The Figure 2 shows the block diagram of the proposed system, the main building blocks are Data Preprocessing Module and Classifier/Classification Module.



Attribute 1	Attribute 2	Attribute n	Class/Items

Attribute 1	Attribute 2	Attribute n	Class/Items

2.3 Block Diagram of the Proposed System

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

It includes

Data Integrationincludes collecting data from various sources. The collected data may be raw in nature. This raw data is preprocessed before using.

String to Integer conversionis performed so that all the values will become numeric.

This conversion can be done using Mapping or Introducing new columns(features) in dataset. Mapping uses key-Value pairs. For example, the color feature contains data {Red, Blue, Green, Yellow, Red} then we can map these colors to values {1, 2, 3, 4, 1} respectively. The drawback of Mapping is that it establishes an Ordinal Relationship between variables i.e. (Red, 1) < (Blue, 2) < (Green, 3) < (Yellow, 4) which is not true.

In the second way, the new columns are introduced. For example, new columns Column_Red, Column_Blue, Column_Green, Column_Yellow are introduced for data {Red, Blue, Green, Yellow, Red} then we fill these columns by appropriate values i.e. 0 or 1. For the given data, Column_Red will have values {1, 0, 0, 0, 1}. Column_Blue will have values {0, 1, 0, 0, 0}, and so on.

Data Normalization aka Feature scaling is a method used to standardize the range of independent variables or features of data.

Min-Max Normalization - This is a simple normalization technique in which we fit the data, in a pre-defined boundary, or to be more specific, a pre-defined interval [A, B] by using the formula,

$$Normalized\ Value = \frac{Value - \ Minimum\ Value}{Maximum\ Value - Minimum\ Value} * (B - A) + A$$

Z Score Normalization aka Standard Deviation method - In this method, the data is normalized by using the formula,

$$Normalized\ Value = \frac{Value - Mean}{StandardDeviation}$$

Feature Selection is the process of selecting a subset of relevant features for use in model construction.

PCA(Principal Components Analysis)can be used for feature selection. It includes following steps.

Step 1: Get data in a Matrix form (say ‘A’).

Step 2: Calculate the mean for each column of ‘A’ and subtract mean from each value of column to derive a new Matrix ‘B’.

Step 3: Calculate the covariance matrix from ‘B’.

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix by using methods like Jacobi.

Step 5: From the eigenvalues and eigenvectors decide features which are essential.
Step 6: Deriving the new data set.

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known. For our system, we have used the K Nearest Neighbor classifier (KNN). It includes following steps.
Step 1: Load the train and test data.
Step 2: Initialize the value of k according to N for Top N recommendations.
Step 3: For getting the top N recommendation, iterate from 1 to total number of training data points
Calculate the distance between test data and each row of training data using different similarity metrics.
Sort the calculated distances in ascending order based on distance values.
For our Top N recommendations, return top k unique values from the sorted array.

Similarity metricis used to quantify the similarity between two objects. Consider, A and B are two vectors of same size. And we have to find the similarity between them. We can use different similarity metrics as follows,
Cosine Similarity- A measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. It uses formula,

$$\boldsymbol{Cosine\ Similarity} = \cos(\theta) = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\,\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Euclidean Similarity- A measure of similarity between two vectors that measures the distance between two respective points of vectors by calculating the length of the path connecting them.The Pythagorean theorem gives this distance between two points.

$$\boldsymbol{Euclidean\ Distance} = d(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Pearson Correlation Coefficient- A measure of similarity between two vectors that quantifies how well two data objects fit a line.

$$\boldsymbol{Pearson\ Correlation\ Coefficient} = \frac{\sum_{i=1}^{n} A_i B_i - \frac{\sum_{i=1}^{n} A_i \sum_{i=1}^{n} B_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} A_i^2 - \frac{\left(\sum_{i=1}^{n} A_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} B_i^2 - \frac{\left(\sum_{i=1}^{n} B_i\right)^2}{n}\right)}}$$

The operation of channel separation is applied on the watermarked color image to generate its subimages, and then 2-level discrete wavelet transform is applied on the subimages to generate the approximate coefficients and detail coefficients.

3. EXPERIMENT AND RESULT

To test our module, we have chosen the wine data set which is available the internet [1].The system is developed in Java programming language using Weka libraries. The PC for experiment is equipped with an Intel Core 2 Duo 3 GHz CPU and 4 GB RAM.
The proposed scheme is tested using the Cross-Validation technique with 10 folds. From the simulation of the experiment results, we can draw to the conclusion that this method is robust.

Table -1 Data Set Information

Data Set Name	Number of Attributes	Number of Instances	Number of Classes
The Wine Data set	13	178	3

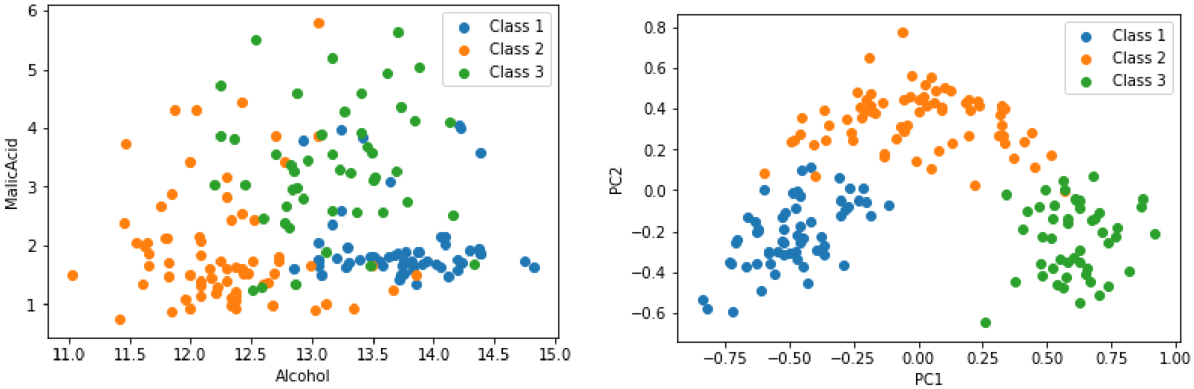
The wine data set are results of a chemical analysis of wines grown in the same region in Italy but,derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Table -2The Attributes

Attribute Name	Type
Alcohol	Numeric
Malic acid	Numeric
Ash	Numeric
Alcalinity of ash	Numeric
Magnesium	Numeric
Total phenols	Numeric
Flavanoids	Numeric
Nonflavanoid phenols	Numeric
Proanthocyanins	Numeric

Color intensity	Numeric
Hue	Numeric
OD280/OD315 of diluted wines	Numeric
Proline	Numeric

The system is evaluated by using different evaluation metrics.
Recall = True Positive / (True Positive + False Negative)
Precision = True Positive / (True Positive + False Positive)
Classification Accuracy = (Items Correctly Classified/ Total Number of Items) X 100 = (True Positive + False Negative)/ (True Positive + True Negative + False Positive + False Negative)\
Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)
Confusion Matrix



(a) (b)

Scatter Plot (a) Before Data Pre-processing (b) After Data Pre-processing

Table -3 Experiment Result

Correctly Classified Instances	169	94.9438 %
Incorrectly Classified Instances	9	5.0562 %
Kappa statistic	0.9238	
K&B Relative Info Score	16414.1043 %	
K&B Information Score	257.5086 bits	1.4467 bits/instance
Class complexity | order 0	278.9816 bits	1.5673 bits/instance
Class complexity | scheme	69.1539 bits	0.3885 bits/instance
Complexity improvement (Sf)	209.8278 bits	1.1788 bits/instance
Mean absolute error	0.0413	
Root mean squared error	0.1821	
Relative absolute error	9.3973 %	
Root relative squared error	38.8682 %	
Total Number of Instances	178	

Table -4 Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.042	0.922	1.000	0.959	0.940	0.983	0.936	Class 1
0.873	0.000	1.000	0.873	0.932	0.897	0.941	0.929	Class 2
1.000	0.031	0.923	1.000	0.960	0.946	0.983	0.917	Class 3
0.949	0.022	0.953	0.949	0.949	0.925	0.966	0.928	Weighted Average

Table -5Confusion Matrix

a	b	c	<- Classified as
59	0	0	a = Class 1
5	62	4	b = Class 2

0	0	48	c = Class 3

4. CONCLUSION

From the experiment results, we can conclude that our system is robust and works fine for the given data set. It generates top N recommendations by specifying the value of k in KNN algorithm. The data pre-processing is necessary to increase the accuracy of prediction.

5. REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/Wine

[2] http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[3] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), “Recommender Systems Handbook”, ISBN 978-0-387-85820-3, Springer, 2011.

[4] C. Yin and Q. Peng, “A Careful Assessment of RecommendationAlgorithms Related to Dimension Reduction Techniques,”Knowledge-Based Systems, Vol. 27, pp. 407-423, 2012.

[5] B. Sarwar, G. Karypis, J. Konstan and J. Reidl, “Analysis ofRecommendation Algorithm for e-Commerce,” in Proceeding ofthe ACM Conference on Electronic Commerce, pp. 158-167, 2000

[6] M. Pazzani, “A Framework for collaborative, content-based anddemographic filtering,” Artificial Intelligence Review, Vol. 13, No.5-6, pp. 393-408, 1999.

[7] D. Jannach, M. Zanker, A. Felfering, G. Friedrish, "RecommenderSystems An Introduction," Combridge university press, 1st edition,2011.

[8] B. Sarwar, G.KaKarypis, J. Konstan and J. Reidl, “Application ofDimmensionality Reduction in Recommender Systems: A CaseStudy,” in proceeding of the WebKDD workshop at the ACMSigKDD, 2000.

[9] Nazila Panahi, Mahrokh G. Shayesteh, Sara Mihandoost, Behrooz Zali Varghahan, “Recognition of Different Datasets Using PCA, LDA, and Various Classifiers,” ISBN 978-1-61284-832-7, IEEE, 2011

[10] Salvador García, Julián Luengo, Francisco Herrera, “Data Preprocessing in Data Mining,” ISBN 978-3-319-10246-7, Springer International Publishing Switzerland 2015