Comparative Study of Recommendation Algorithms and Systems using WEKA

Lokesh S. Katore
Student
Computer Engineering
Pmpri Chnchwad College of Engineering

J.S.Umale, Ph.D.
Professor
Computer Engineering
Pmpri Chnchwad College of Engneering

ABSTRACT

Recommendation systems now days are the heart of success stories for business and optimization of resources. The accurate prediction of business decision accurately depends on heuristic algorithms used for analytics. Classical algorithms used for the data mining find their utility to perform with the new challenges considering key factors for improvement. This paper presents the performance of the specific algorithms of the data mining class in view to observe their suitability for recommender systems.

Keywords

Data mining, WEKA, J48, Naïve Bayes, Simple Cart, K Star, Resample, SMOTE.

1. INTRODUCTION

Data Mining [1] is sometimes called as data or knowledge discovery. It is the way or process of analyzing data from different perspectives and summarizing it into useful information. Data mining is the process of finding frequently occurring patterns or correlation among different fields in large relational database. Recommender System means the system which is specially meant for predicting or recommending. In earlier days fortune teller use to tell the fortune with the help of crystal ball. Meteorologist uses maps and scientific data to tell us about the possibility of rain, snow or sunshine. Manual techniques of prediction sometimes give accurate prediction and sometimes get miserably failed. Numbers of things are planned according to the prediction. Failure of prediction leads to loss. So to have accurate decision every time with fewer flaws; Automation of recommender system is required essentially. To recommend there are several algorithms available in data mining namely Cart, J48, Simple regression, Apriori, FP growth, K Star, Naïve Bayes etc. To make the system automated one need to know which algorithm is best suitable for the dataset. To know the performance of the algorithm accuracy is to be measured. The algorithm showing the best accuracy is most suitable to be used for predicting or recommending.

Some dataset are noisy, inconsistent and pertaining missing values. These datasets need to be preprocessed. Preprocessing means cleaning, removing inconsistencies, filling the missing values in the dataset. This paper comprises the knowledge level testing data collected from UCI repository. On this dataset four algorithms are applied to check their accuracy. On the basis of accuracy algorithms performance is compared. This paper mainly focuses to increase the accuracy of the algorithm with the help of filter namely smote and resample.

WEKA [3] is a collection of machine learning algorithms for data mining. The algorithms can be applied directly to a dataset. WEKA contains tools for data classification, Association, clustering. This paper focuses on algorithms like Naive Bayes, J48, Simple Cart and K star. WEKA is used for pre-processing and performance comparisons. The feature selection in the tool describes the attribute status of the data present in Knowledge level dataset.

2. EXPECTED CHARACTERISTICS REQUIRED FOR RECOMMENDATION ALGORITHM

This paper focuses on the following three measures namely correctly classified instances, incorrectly classified instances, accuracy [1].

(i) Correctly classified instance:
These are the instances which are correctly classified by any classification algorithm. Percentage of correctly classified instances is called as accuracy.

(ii) Incorrectly classified instances:
These instances are not correctly classified by the algorithm. Sometimes it is observed that the data which is incorrectly classified may contain inconsistent data, noisy data or data out of scope.

(iii) Accuracy:
Accuracy is how a measured value is close to the true value. The general formula is given below:

$$\text{Accuracy} = \frac{Tp+Tn}{P+N} \text{------- (1).}$$

In equation 1,

Tp indicates True positive, Tn indicates True negative, P indicates Total positive, N indicates Total negative.

Where, $P = Tp + Fn$, $N = Fp + Tn$.

Accuracy is an important factor to analyze the performance of an algorithm. Accuracy is the ratio of sum of true positive value and true negative value to the total positive and total negative value. In recommender system, the algorithm with highest accuracy will be selected for the recommendation. Accuracy of the algorithm varies according to the dataset used. So before using the algorithms for recommender system, we must check the accuracy of the algorithm. So it will reduce the cost of doing trial and error of using algorithms in Recommender system. The Algorithms used for analysis is Naïve Bayes, K Star, J48 (C4.5), Simple Cart.

3. METHODS AND ALGORITHMS

3.1 Naïve Bayes [1][9]

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem [1]. It is a probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. Diagnostic and predictive problems can be solved. Bayesian classification provides practical learning algorithms and prior knowledge to the observed data. Bayesian Classification is used for understanding and evaluating many learning algorithms. An explicit probability for hypothesis is calculated by the Naïve Bayes and it is robust to noise in input data.

The main drawback of this approach is it is feature based. It only checks the presence and absence of the feature[9]. Sometimes it may predict wrong things as it does not consider other feature. For example, red, round is a feature of apple and another fruit is red and round, so it will consider it as an apple, but it might be the another fruit.

3.2 K Star [4][11]

K star is a nearest neighbor method with generalized based on transformations. It is lazy learning classification. It provides a consistent approach to handle symbolic attributes, real valued attributes and missing values [4]. It is lazy learning approach specially meant for cluster analysis. It uses entropic distance measure for prediction. Space required for the storage is very large as compared to other algorithms. Mostly noisy training data increases the case support unnecessary. It is usually slower in evaluating the result[12].

3.3 J48 [5]

J48 classifier is a simple C4.5 decision tree for classification. It is supervised method of classification. It creates a small binary tree. It is univarient decision tree. It is an extension of ID3 algorithm. In this Divide and Conquer approach is used to classify the data. It divides the data into range based on the attribute value for that value that are found in training sample. As this approach is range based and univarient[11], it does not perform better than multivarient approach. As this is decision tree approach it is very much useful in predicting the values. J48 accuracy of correctly classified instance is much more than that of the other algorithms, are univarient in nature[10].

3.4 Simple Cart [9]

Classification and Regression Trees is a classification method which uses historical data to construct decision trees. Decision trees are used for the classification of new data. In order to use CART we need to know number of classes a priori [9]. CART uses learning sample to build a decision tree. Sample is a set of historical data with pre-assigned classes for all observations. Take example of learning sample for credit scoring system, the sample would be basic information about previous borrows matched with actual payoff results [9].

3.5 Filtering [6]

The very simple techniques like filtering are fruitful in increasing the accuracy of a result shown by data mining algorithms. Here two filters are applied to make imbalance data balance, they are SMOTE and Resampling.

SMOTE is an instance filter used in supervised learning. It is mainly used to balance the data sets. Balance data sets are those data sets which have approximately as many as positive example of the concept as there is negative ones. There exist many domains which do not have balance data sets. The problem with imbalance class is that the standard learners are often biased towards majority class, because these classifiers attempt to reduce global quantities such as the error rate without taking the data distribution into consideration. As the result example from the overwhelming class are well classified whereas examples from minority class tend to be misclassified. SMOTE [6] informed oversampling generalize the decision region for the minority class. As a result, larger and less specific regions are learnt, thus paying attention to the minority class samples without causing over fitting. Resampling can be done with or without replacement of the sample data to produces a random subsample. The original dataset must fit into entire space present in the memory. The number of instances present in the generated dataset may be specified. The dataset must have a nominal class attribute [6]. This filtering is also used for balancing the imbalance dataset.

4. EXPERIMENTAL SETUP

Algorithms selected are namely J48, Naive Bayes, Simple Cart and K star. These algorithms are to be compared on the basis of the accuracy, when applied on knowledge level dataset. A 10 fold cross validation is used for validating the results. The data mining method used to build the model is classification. The training data set consists of 258 instances and 7 attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of Knowledge level data. The performance of the classifiers is evaluated and their results are analyzed. The results of comparison are based on 10 ten-fold cross-validations.

4.1 Work Flow

The Figure 1 is showing the process of how to apply algorithms on data sets with and without filtering. To use Smote and Resample as a filter techniques, the procedure is same but only difference is before applying classification algorithm SMOTE and Resample should be applied for filtering the data. Filtering is mainly used to avoid over fitting of the data.

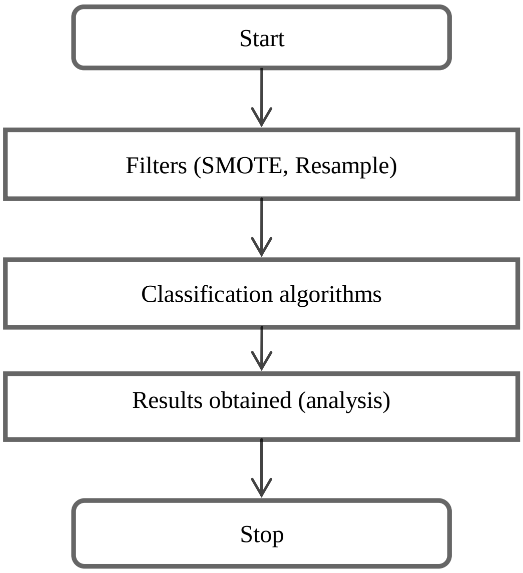
Start
Filters (SMOTE, Resample)
Classification algorithms
Results obtained (analysis)
Stop

Fig 1: Implementation of Algorithms for accuracy analysis using WEKA.

Table 1. Analysis of algorithm on Knowledge level Dataset

Classification Techniques	Correctly Classified Instance	Incorrectly Classified instance	Kappa Statistic	Mean absolute error	Root mean squared error	Relative Absolute error	Root relative squared error	Total Instance	Accuracy
C4.5(J48)	241	17	0.907	0.0404	0.1751	11.32%	41.506%	258	93.41%
Naïve Bayes	230	28	0.8471	0.111	0.2352	31.14%	55.74%	258	89.14%
Simple Cart	236	22	0.8793	0.0634	0.1948	17.78%	46.15%	258	91.47%
K Star	203	55	0.6957	0.1127	0.2668	31.61%	63.22%	258	78.68%

5. RESULTS AND DISCUSSION

The experimentation on the knowledge level dataset has been done in order to check the performance on the basis of accuracy. Out of four algorithms C4.5 (J48) is showing highest accuracy i.e. 93.41% than the other three. It depicts that C4.5 is performing better and will give good prediction results than the other three algorithms for the Knowledge level dataset consisting of 258 instances and 6 attributes. From the Table 1 the algorithms namely Naïve Bayes, K Star and Simple Cart have 89.14%, 91.47% and 78.68% accuracy respectively. The 93.41% accuracy means our algorithm will give 93.41% good results. 7% incorrect result will be displayed which is much better than the other algorithms which are showing 11%, 9% and 22% incorrect results respectively. The results are obtained before applying the filters to avoid over fitting of the data. After applying filters on the dataset; the accuracy of the algorithms is being increased.

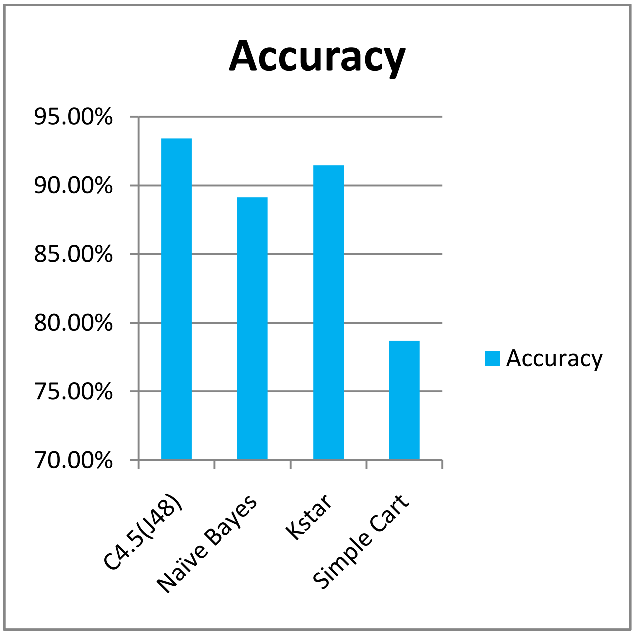

Fig 2. Comparisons of Accuracies before applying filters to the dataset

Figure 2 shows the accuracy of the algorithms in terms of percentage. Here, we can clearly observe that the C4.5 algorithm accuracy is more as compared to other algorithm.

Accuracy means the ratio of true positive values and true negative value to the positive and negative value.

Table2. Analysis of accuracy of algorithms after using filters (SMOTE & Resample).

Classification Techniques	Accuracy	Accuracy after filtering (SMOTE & Resample)
C4.5(J48)	93.41%	97.51%
Naïve Bayes	89.14%	89.004%
Simple Cart	91.47%	93.61%
K Star	78.68%	91.48%
Average	88.18%	92.90%

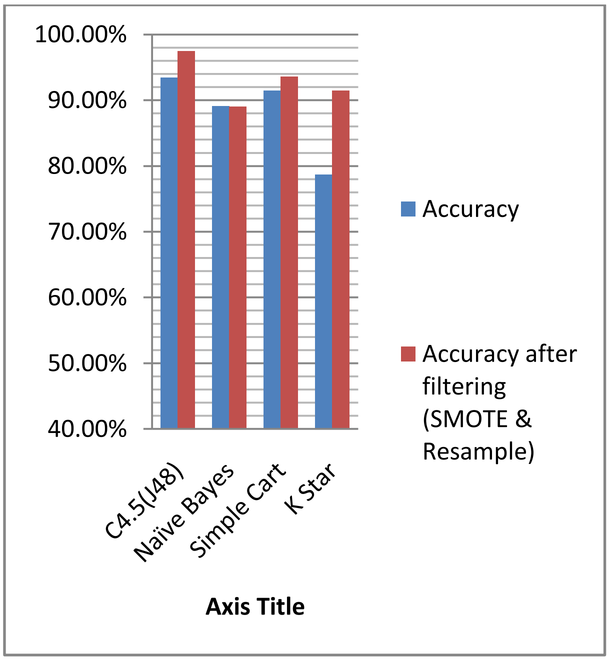

Fig 3. Accuracy before and after filtering

Accuracy of all the algorithms increased after using filter as a preprocessing technique. Average percentage increase in the accuracy is 10.53%. The maximum increase in the accuracy is being seen in K Star algorithm i.e. is 11.62%.

6. CONCLUSION AND FUTURE WORK

In this paper, Experimentation on Knowledge level dataset is being performed using four algorithms namely C4.5, Naïve Bayes, K Star and Simple Cart and compared according to their accuracies. As we have seen the C4.5 algorithm performance better than the other for the 258 instance and 6 attribute dataset. We have analyzed that after using the Filter (SMOTE and Resample the accuracies of the algorithms is being increased. So to have a proper recommend der system with less flaws accuracy must be compared.

7. REFERENCES

[1] Naive Bayes text classification (http://nlp.stanford.edu/IR-book/html/htmledition/naive bayes-text-classification-1.html) accessed on March 14.Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[2] Dataset of Liver, Bank and dermatology (www.repository.seasr.org/Datasets/UCI/arff/) accessed on March 14.

[3] Information about Weka and its application (www.cs.waikato.ac.nz/ml/weka).

[4] K*: An Instance-based Learner Using an Entropic Distance Measure Authors John G. Cleary, Leonard E. Trigg, Dept. of Computer Science, University of Waikato,NewZealand.email:{jcleary,trigg}@waikato.ac. nz.

[5] Tina R. Patil, Mrs. S. S. Sherekar ,Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification

[6] B Imbalanced Data Set Learning with Synthetic Examples, Authors Benjamin X. Wang and Nathalie Japkowicz.

[7] University of Waikato,(www.waikato.ac.nz.) accessed on March 14.

[8] S Analysis of Liver Disorder Using Data mining Algorithm, Global journal of computer science and technology, Vol. 10 Issue 14 (Ver. 1.0) November 2010 Page | 48

[9] https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Naive_Bayes_classifier.html naive bayes.

[10] Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava, Manish Mathuria; International Journal of Advanced Research in Computer Science and Software Engineering

[11] Ms S. Vijayarani ,Ms M. Muthulakshmi; Comparative Analysis of Bayes and Lazy Classification Algorithms.

[12] Sanidhya Painuli,M. Elangovan, V. Sugumaran; Tool condition monitoring using K-star algorithm