A CONTEMPORARY LOOM FOR CUSTOMIZATION AND VISUALIZATION KNN AND NAIVE BAYES ALGORITHMS

BIPIN NAIR B J
Lecturer in Department of Computer Science, Amrita Vishwa Vidyapeetham, Mysore Campus, Karnataka, India
bipin.bj.nair@gmail.com

ABSTRACT: The purpose of the paper "A Contemporary Loom For Customization and Visualization of KNN and Naive Data Mining Algorithms" aims at customizing the existing data mining algorithms to provide an interpretable output with an efficient and interactive visualization. The Data mining algorithms are chosen from specific areas like classification. The algorithms are made customizable such that it can adapt to user requirements. The visualization provided for the mined outputs are made interactive such that even a naïve user can interpret the output.The available data mining tools require expert users to carry out experiments. The scope of this relies in the area of visualization of the mined data. The visualizations can be made interactive such that the outputs can be easily interpreted by the user with less effort.

KEYWORDS: Data Mining (DM)); Attribute Relation File Format (ARFF):k-neighbour and nearest(KNN)

INTRODUCTION

Data Mining is the process of analysing data from different prospective and summarizing it into useful information .Data Mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it and summarize the relationship identified. Technically Data Mining is the process of finding correlations or patterns among dozens of fields in a large relational database. Data mining often involves the analysis of data stored in a data warehouse. One of the major data mining techniques is classification. Data classification is the categorization of data for its most effective and efficient use. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Popular classification techniques include KNN and naive bayse. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits.

OVERVIEW

This paper deals with the analysis, customization and visualization of data mining algorithms for various datasets. The system provides with choice of algorithms and its respective customizations that can be done by the user. The results are created such that they can be easily interpreted. The outcomes are also given a visual impact with an interactive environment so that it can be interpreted easily.

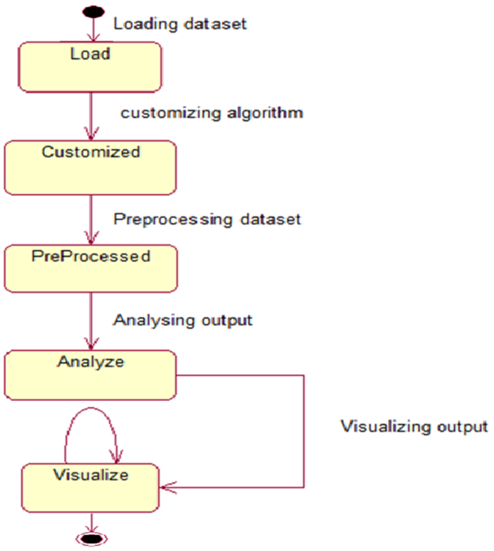

Figure 1. Working of Algorithm

PROBLEM STATEMENT

There are many data mining tools available today. There are excellent tools that are available. But they lack in some points. They are,

- Sometimes expert support is required to use these tools.
- The dataset should have a common format. If not, it will ask for the conversion to its own format. The naive user finds difficulty to do this step.
- They provide limited user interaction.
- The processed output of these tools is not completely understandable by a naive user. It provides some terms which needs an expert's help to understand

PROBLEM FORMULATION

The proposed paper helps the user to work with the major data mining algorithms in an easy way, and the user is provided with an interactive and effective visualization that helps to make a useful decision. It mainly overcomes the limitations of the existing work in terms of user involvement and visualization.

The Proposed work performs,

- Developing an understanding of the application domain, relevant prior knowledge and the goals of the end-user.
- Selecting a dataset on which discovery is to be performed.
- Data Pre-processing: This stage includes operations for dimension reduction (such as feature selection and sampling); data cleansing (such as handling missing values, removal of noise or

outliers); and data transformation (such as discretization of numerical attributes and attribute extraction).

- Choosing the appropriate data mining task such as classification, regression, clustering and summarization.
- Choosing the data mining algorithm. This stage includes selecting the specific method to be used for searching patterns.
- Employing the data mining algorithm.
- Evaluating and interpreting the mined patterns

RELATED WORK

- Visualization Module
- Request Handler

The request handler consists of a controller class which controls the requests and responses. Controllers provide access to the application behavior which is typically defined by a service interface .Controllers interpret user input and transforms such inputs into a sensible model which will be represented to the user by the view. Controller is a single method that is responsible for handling a request and retrieving an appropriate model and view. The DB Controller controls the relational data store .The details of the user are stored in a local data base
Each Algorithm has tree formatter.

- Input
- Output
- Visualization

Naive bayes Algorithm

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Byes' theorem can be stated as follows.
Naive Bayes Pseudocode :

Input : A data Set, and a training instance.

Output : The class in which training instance belongs to.

Methods :

1)Create a Naive Bays class ,
2)

- Read the input data from an external file(Training data)
- Copy each data instance to be an n-dimensional vector of attribute Values, **X = (x1 , x2 , x3 , . . . , xn),** where n=number of attributes.
- Read the testing data for the classification
- Call the classification method in Classifier class.

2)Create Classifier class
Do training as:

In a classifier which assigns the test instance to one of m classes C1, C2..., Cm. A data instance X is assigned to the class for which it has the highest posterior probability conditioned on X, i.e. the class which is most probable given the prior probabilities of the classes and the data X.

```
IF $P (Ci |X) > P (Cj |X)$ ,for all j such that $1 \leq j \leq n, j = i$.

THEN

    X is assigned to class Ci.

ENDIF
```

Now calculate

$P (Ci |X)=(P (X|Ci) P (Ci)) /p(X)$

Since P (X) is a normalising factor which is equal for all classes, we need only maximise the numerator ,P (X|Ci)P (Ci) Estimate both the values , P (X|Ci) and P (Ci),

Estimating the class prior probabilities

P(C1)=(no of occurrence of outcome C1)/(Total no of occurrences);

P(C2)=(no of occurrence of outcome C2)/(Total no of occurrences);

Calculate the probability of each attribute for each class.

Repeat this for every class **C1,..Cn**

P (X = (Attribute_1, attribute_2,..., Attribute_n)|C1) = P (attribute_1|C1) × P (attribute_2|C1) ×..... x P (attribute_3|C1)

put this together with our known prior probability for class C1 to obtain,

P (C1 |X = ((Attribute_1, attribute_2,..., Attribute_n)) =P (X = (Attribute_1, attribute_2,..., Attribute_n)|C1)P (C1)) / p(X)

Now calculate it for the other class

P(X = (Attribute_1, attribute_2,..., Attribute_n)|C2) = P (attribute_1|C2) × P (attribute_2|C2) ×.....x P (attribute_n|C2) put this together with our known prior probability for each class to obtain,

Repeat this for Cm classes .

P(C1 |X = (Attribute_1, attribute_2,..., Attribute_n)) =(P (X = (Attribute_1, attribute_2,..., Attribute_n)|C1)P (C2)) / p(X) after finding the probabilities of p(C1/X) , p(C2/X),.... ,p(Cm/X)

```
IF  p(C1/X) > p(C2/X)

THEN
        Assign the new instance to class C1
ELSE
        Assign the new instance to class C2.
END
```

KNN Algorithm

K-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. The classification rules are generated by the training samples themselves without any additional data. The KNN classification algorithm predicts the test sample's category according to the K training samples which are the nearest neighbours to the test sample, and judge it to that category which has the largest category probability.

KNN Pseudo code :

```
Begin
Read the training data from a file S

Read the testing data from a file  x`

Consider k as the desired number of nearest neighbors and $S: = p_1,...,p_n$ be the set  of training
samples in the form $p_1 = (x_i,c_i)$, where $x_i$ is the d-dimensional feature vector of the point $p_i$ and $c_i$
is the class that $p_i$ belongs to.

Calculate k,  where k = n^(1/2).

For each p'=(x',c')

●   Compute the distance $d(x',x_i)$ between $p'$ and all $p_i$ belonging to
●   Sort all points $p_i$ according to the key $d(x',x_i)$
●   Select the first $k$ points from the sorted list, those are  the $k$   closest training samples to $p'$
●   Assign a class to p' based on majority vote
```

EXPERIMENT RESULT

Naive Bayes

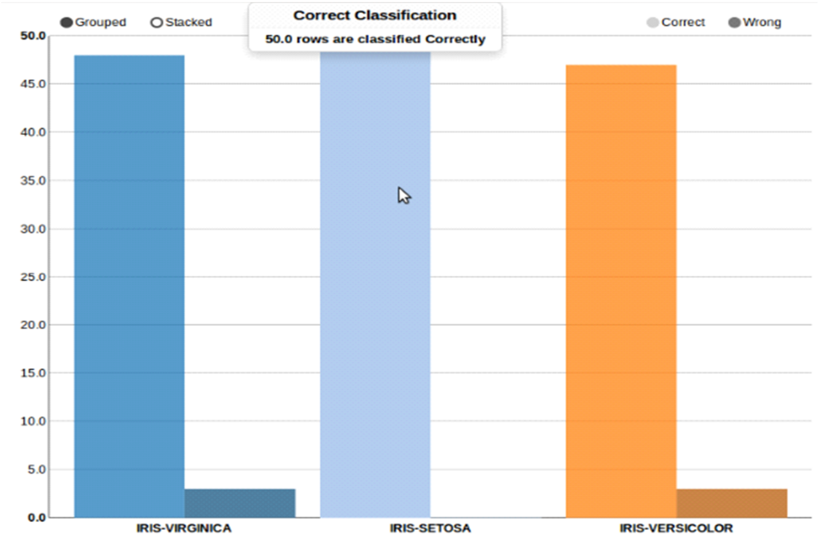

Fig 2 .Grouped Bar Chart for iris DataSet with tooltip

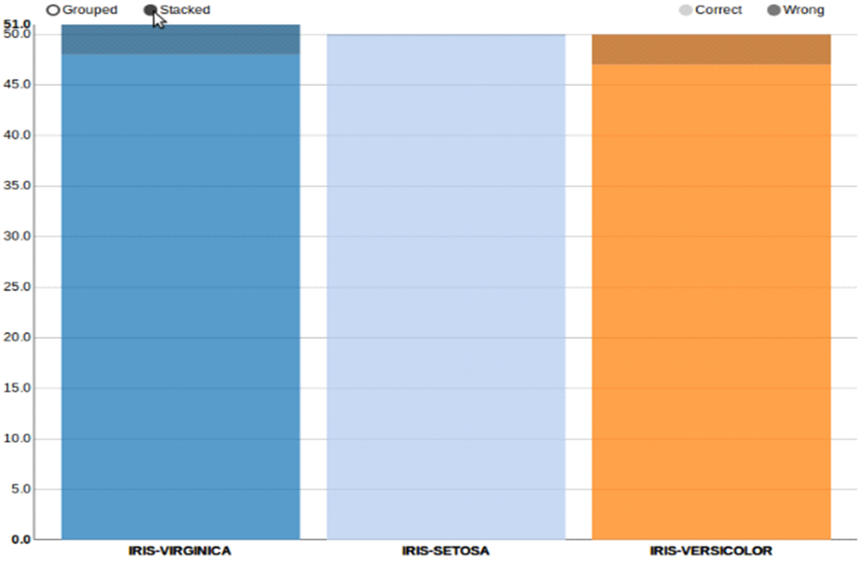

Fig 3.Stacked Bar Chart for iris Data Set

KNN

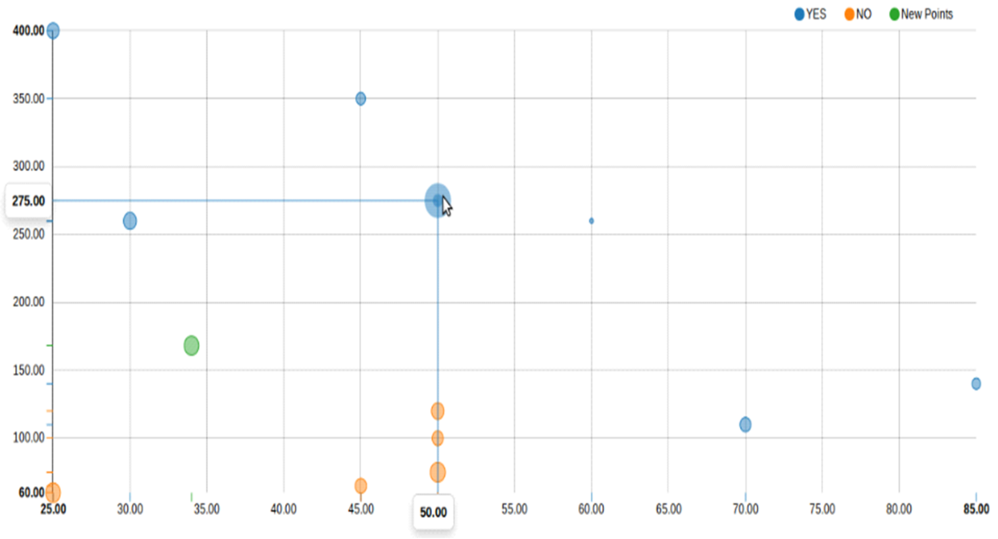

Fig 4. Scatter chart for xy data set in KNN

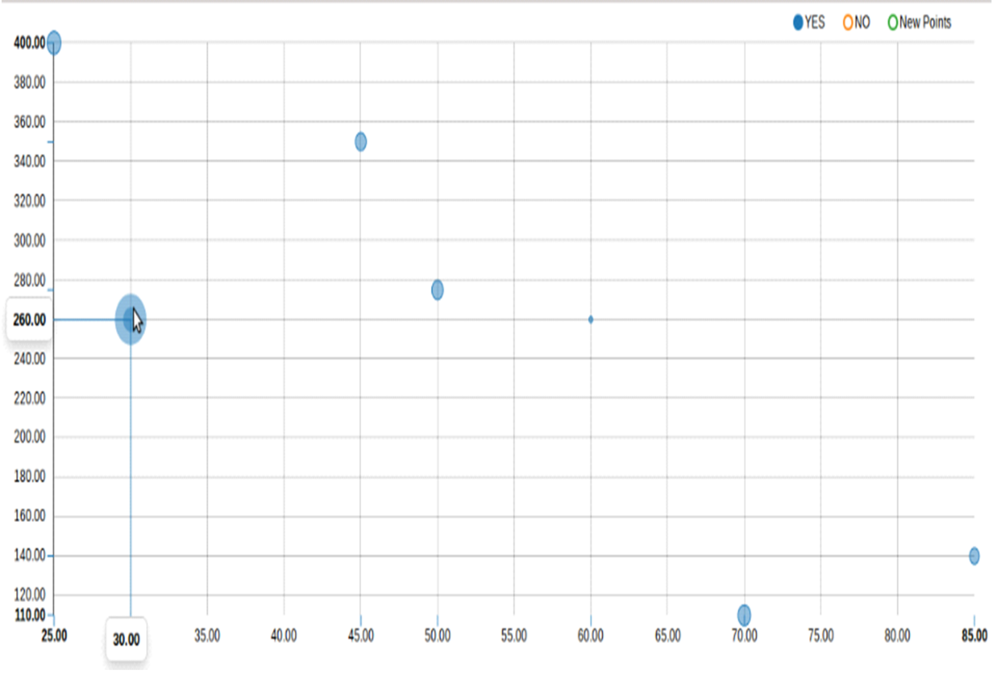

Fig 5. Data of class YES in XY dataset

CONCLUSION

The paper Analysis, Customization, and Visualization of KNN and naive bayes Data Mining Algorithm, had been started as a stand-alone application .The technology used was inefficient to adapt to the specified requirements. Hence the paper has been changed from a stand –alone application to web based application which uses spring framework for the integration and d3.js for the visualization.

This work is part of a research paper, so there were time constraints in the implementation of the paper. As the technologies being used for this were changed in order to come up with the best result, all that has been achieved within this time is the final paper.

REFERENCES

Data Mining- Introductory and Advanced Topics

Data Mining: Concepts and Techniques by Jiawei Han

A Comparative Analysis of Data Mining Tools in Agent Based Systems by Sharon Christa, K. Lakshmi Madhuri, V. Suma

CSE5230 Tutorial: The Na ̈ve Bayes Classifier

The Apriori Algorithm – a Tutorial by Markus Hegland

DATA MINING: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.

Naive Bayes Classifier example - Eric Meisner lecture05-NaiveBayes-2up.pdfAn Improved k-Nearest Neighbor Classification Using Genetic Algorithm ,N. Suguna1, and Dr. K. Thanushkodi2

International Journal of Computer Trends and Technology (IJCTT) - volume4Issue4 –April 2013 ENHANCED DBSCAN ALGORITHM by Priyamvada Paliwal#1, Meghna Sharma.

An Introduction to Neural Networks by Vincent Cheung and Kevin Cannons.

Jiawei Han and Micheline Kamber, University of lllinois at urbana-champaign, Concepts and Techniques, Data mining, second edition(2006), [3]Barandela, R., Sánchez, J.S., García, V., Rangel, E. Strategies for Learning in Class Imbalance Problems. Pattern Recognition 2003, 36(3), pp.849-851

K.P Soman, Shyam Diwakar, V.Ajay, data mining theory and practice, Amrita Vishwa Vidyapeetham, (2010).

Data Mining: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.

Apriori Algorithm Review for Finals Presentation by SE 157B, Spring Semester 2007 Professor Lee By Gaurang Negandhi.

PDF-”Using TF-IDF to Determine Word Relevance in Document Queries” by Juan Ramos, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855

Figure 1 – Schematic drawing made through Rational Rose.

Figure 2 to 5- Resulting Snapshots obtained by implementing algorithms Naïve Bayes and KNN.