Attribute and value extraction in classification in data mining

Prof. Sharayu G. Karandikar
Lecturer
Dr. G.D. POl Foundation, YMT College of Management
MCA Department
Sgk_010606@yahoo.co.in

Abstract--Every algorithm in data mining requires precise, authentic and well prepared data, to produce useful and accurate results. Valuable results produced by data mining will deliver knowledge that will be used in complex analysis and decision support. Data preparation in data mining includes extraction of useful attributes from database, setting up appropriate values for the selected attributes and preparing training and sample data sets for implementation of the algorithm. This process is lengthy, time consuming, tedious and equally important. A methodology should be adopted for setting up values for selected attributes in a dataset. This paper presents a method for setting up attributes with derived values. It also mentions significance of metadata to record the steps followed for deriving values. Then a practical approach is adopted to demonstrate the method implementation on sample data in Naïve Bayes classification algorithm in data mining. New features for future enhancement are suggested.[2]

Key words: data mining, attribute, derived value, classification, algorithm, Naïve Bayes theorem

INTRODUCTION

Data mining techniques are implemented to a single table or view under consideration. The actual data resides in multiple tables and views. The attributes selected for implementation of data mining algorithm are as per given problem definition. The attributes from various data objects are collectively stored in a single dataset. For some of the selected attributes values are imported as it is to the dataset. For some attributes, new values are derived as per the requirement of algorithm under consideration. This paper presents a methodology for attribute and value extraction that involves following steps [1]

Attribute selection

Selection of appropriate attribute is based on the problem definition. The objective of this step is to include all dependent as well as independent attributes necessary to generate required results. The attributes are selected from various data sources and belongs to various data types. In some cases related attributes can be combined or a single attribute can be split into multiple attributes as per requirement.

Value extraction

After attributes are finalized the values under every attribute is to be extracted from the data source. Here we have considered built in data types like logical, numeric and non descriptive text. Values pertaining to logical data type are discrete and form a finite set. Even then, if required these values can be grouped together to form specific number of classes. Data that belongs to numeric type may be divided into ranges for generating classes. Text data can be trimmed or modified as per requirement of class generation.

Data labeling

The classes formed for every individual attribute needs to be labeled correctly. Appropriate data labeling techniques are adopted so that the class label can represent set of values accurately. The data stored under every attribute may be trimmed and tagged to fit in the specified class. The outliers are identified and taken care off. Now the data is fully prepared and ready to undergo implementation of specified algorithm.

Metadata

It is extremely important to maintain details of every step in the process. Attributes and original values undergo many changes in the process of class formation. All these changes should be recorded systematically so that it can be retrieved and used in future by new data items or in reverse engineering to retrieve back original set of attributes and values. Further this paper presents practical working out of above methodology on a sample dataset for Naïve Bayes classification algorithm in data mining.

Outline

The remaining sections of the paper are organized as follows :

Section 2: Related work is described.
Section3:Research methodology is presented
Section4:graphical representation of methodology
Section5:proposed methodology is applied for Naïve Bayes classification Algorithm
Section6: Conclusion and suggestions

RELATED WORK

Data preparation is a repetitive process and major contribution in data preparation is of attribute selection and value extraction. Some researchers have taken a step ahead to make it systematic and methodical. This work can be mentioned as In the paper "class driven attribute extraction " the author has reported on the large-scale acquisition of class attributes with and without the use of lists of representative instances, as well as the discovery of unary attributes, such as typically expressed in English through pronominal adjectival modification.
In an article "A Self-Supervised Approach for Extraction of Attribute-Value Pairs from Wikipedia Articles" the author has presented a self-supervised approach for autonomously extract attribute-value pairs from Wikipedia articles. We apply our method to the Wikipedia automatic info box generation problem and outperformed a method presented in the literature by 21.92% in precision, 26.86% in recall and 24.29% in F1.
In the paper "Text Mining for Product Attribute Extraction" the author we describe our work on extracting attribute and value pairs from textual product descriptions. The goal is to augment databases of products by representing each product as a set of attribute-value pairs. Such a representation is beneficial for tasks where treating the product as a set of attribute-value pairs is more useful than as an Atomic entity.
The website "www.docs.oracle.com" has provided valuable and detailed information on data mining and all its phases.
In a book "Data Preparation for Data Mining" by Dorian Pyle the author has addressed an issue unfortunately ignored by most authorities on data mining: data preparation.
In the paper "Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions" the author has explained an approach to extract attribute-value pairs from product descriptions. This allows us to represent products as sets of such attribute-value pairs to augment product databases. Such a representation is useful for a variety of tasks where treating a product as a set of attribute-value pairs is more useful than as an atomic entity.

RESEARCH METHODOLOGY [3]

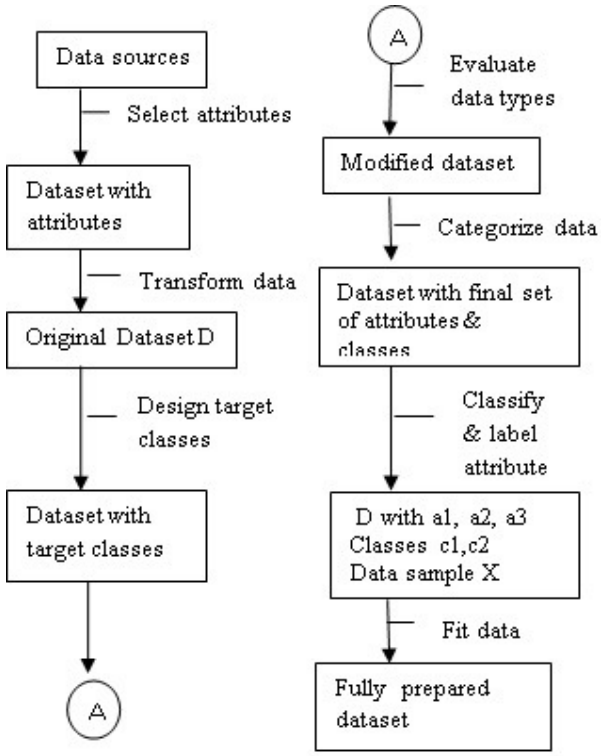
For research purpose a methodology is adopted that is described above. It is with reference to the data preparation step in data mining. This methodology depicts precise steps for deriving data for selected attributes in a sample dataset for Naïve Bayes classification algorithm in data mining.

A.Select attributes as per requirement given in problem statement.
B.Study, requirements of algorithm with respect to data preparation.
C.Consider sample dataset with selected attributes and original values.
D.Evaluating data on the basis of data type.
E.Categorize data under each attribute as discrete, ranged, and categorical.
F.Define classes for each attribute and label them. Also define target classes and label them.
G.Fitting data into classes with the help of trimming and tagging.
H.Execution of each step is recorded and documented as metadata.

GRAPHICAL REPRESENTATION OF THE METHODOLOGY FOR NAÏVE BAYES CLASSIFICATION ALGORITHM [4]

Following diagram shows the steps in the methodology. Metadata preparation is implicit on every step. It will help to track changes in original data as well as record each data process. All the formulae and methods used for computation are stored. Dataset should be updated as the original data sources are updated.

Following diagram represents steps to be followed



APPLYING METHODOLOGY FOR NAÏVE BAYES CLASSIFICATION ALGORITHM [2]

A.Attribute selection

This example is selected for predicting whether a person will buy a laptop or not with selected attributes like age, occupation and income.

B.Study of algorithm

•Classification assigns items to discrete classes and predicts the class to which an item belongs.

•Naïve Bayes algorithm makes prediction which derives the probability of a prediction from the underlying evidence, as observed in the data.

•For a given target value, the distribution of each predictor is independent of the other predictor.

C.Sample dataset

•Selected Attributes are a1,a2,a3

•Original transformed data in dataset D

•Target Classes are designed as

Person will buy computer –y

Person will not buy computer –n

Dataset D			
Age (a1)	**Occupation (a2)**	**Income (a3)**	**Buy laptop (target)**
24	Student	27000	y
19	Student	55000	y
35	Service	18000	n
38	Business	62000	y
31	Student	30000	y
42	Service	36000	y
47	Student	15000	n
44	Business	40000	n
33	Student	45000	y
20	Student	50000	y

D.Evaluate data type

Age : numeric discrete
Occupation : text
Income : numeric discrete

E.Data categorization

Age : below 30 , between 30 and 40, above 30
Occupation : student and professional
Income : less than 20000 is low, between 21000 and 40000 is medium, more than 40000 is high

F.Attribute classification and labeling

Age :< 30 – l , >=30 and < 40 – b, >40 – a

Occupation : student - yes / no

Income :<20000 – l , >=20000 &< 40000 – m,

> 40000 – h

Target classes :Buy computer : yes (c1)/no (c2)

G.Fitting data

Dataset d after attribute and value extraction				
Sr no	**age**	**student**	**income**	**Buy laptop**
1	l	Y	m	y
2	l	Y	h	y
3	b	N	l	n
4	b	N	h	y
5	b	Y	m	y
6	a	N	m	y
7	a	Y	l	n
8	a	N	m	n
9	b	Y	h	Y
10	l	Y	h	Y

++ CODE FOR IMPLEMENTATION OF NAÏVE BAYES CLASSIFICATION ALGORITHM FOR ABOVE EXAMPLE

```
#include<iostream.h>

#include<conio.h>

void main()

{

char age[10]={'l','l','b','b','b','a','a','a','b','l'};

char student[10]={'y','y','n','n','y','n','y','n','y','y'};

char income[10]={'m','h','l','h','m','m','l','m','h','h'};

char status[10]={'y','y','n','y','y','y','n','n','y','y'};

clrscr();

/* training data set prepared with the help of proposed methodology is represented in terms of array*/

charsg='b',ss='y',si='m';

/* data sample X defined as age=35/student=yes/

income=28000 */

/* X is represented with the help of new computed

Attributes age=b/student=y/income=m */

intrc=10;  // total no of objects are 10

floatpa_y=0.0, pa_n=0.0, ps_y=0.0, ps_n=0.0,

pi_y=0.0, pi_n=0.0,px_y=1.0,px_n=1.0;
```

```
for(inti=0;i<10;i++)
{

if(age[i]==sg&& status[i]=='y')       pa_y=pa_y+1;

if(age[i]==sg&& status[i]=='n')       pa_n=pa_n+1;

if(student[i]==ss&& status[i]=='y') ps_y=ps_y+1;

if(student[i]==ss&& status[i]=='n') ps_n=ps_n+1;

if(income[i]==si&& status[i]=='y')  pi_y=pi_y+1;

if(income[i]==si&& status[i]=='n')  pi_n=pi_n+1;
}

            pa_y=pa_y/rc;     pa_n=pa_n/rc;

            ps_y=ps_y/rc;     ps_n=ps_n/rc;

            pi_y=pi_y/rc;     pi_n=pi_n/rc;

cout<<"\n probability for every attribute is calculated independently as per the steps in algorithm for both the classes of buy laptop c1: yes and c2: no \n";

cout<<"\n"<<"probability for age     : y : "<<pa_y<<"\t n : "<<pa_n<<"\n";

cout<<"\n"<<"probability for student : y : "<<ps_y<<"\t n : "<<ps_n<<"\n";

cout<<"\n"<<"probability for income  : y : "<<pi_y<<"\t n : "<<pi_n<<"\n";

px_y=pa_y*ps_y*pi_y;

px_n=pa_n*ps_n*pi_n;

cout<<"\nprobability for X is calculated for each class c1 and c2 \n";

cout<<"\n"<<"probability of X buying laptop : yes "<<px_y<<"\n";

cout<<"probability of X buying laptop : no "<<px_n<<"\n";

if(px_y>px_n) {cout<<"\nX belongs to class c1 i.e. buy laptop  yes\n";}

else {cout<<"\nX belongs to class c2 i.e. buy laptop no\n";}

}
```

The output

probability for every attribute is calculated independently as per the steps in

algorithm for both the classes of buy laptop c1:yes and c2:no

probability for age   : y : 0.3   n : 0.1

probability for student : y : 0.5   n : 0.1

probability for income : y : 0.3   n : 0.1

probability for X is calculated for each class c1 & c2

probability of X buying laptop : yes 0.045

probability of X buying laptop : no 0.001

X belongs to class c1 i.e. buy laptop yes

CONCLUSION AND FUTURE WORK

Data preparation phase primarily includes attribute selection and value extraction to prepare datasets. It is very important to represent the attribute values in simplest possible way. Also attributes should be categorized into sets as per requirement of number of classes. These classes should be properly labeled and described in details as per given methodology. The data extracted from various data sources under each attribute is formulated with the help of trimming and tagging techniques. The techniques should be recorded as metadata.

This research paper proposes a method of carrying all these data preparation tasks with precise and systematic steps. it considers data pertaining to primary data types like character and number. This new approach will provide an authentic and easy way to derive data values to be fit in the dataset.

These fully prepared datasets can be used to implement data mining algorithm. Data mining is a upcoming field in area of artificial intelligence a subject under machine learning. The data is mined with various algorithms to discover & deliver knowledge. The knowledge is useful in analysis and decision making activities in enterprises. [1]

Experts and researchers may take a step ahead to consider various data types. The method can be found which may take care of attributes with discrete, categorical as well as mixed values. Domain experts may go a step further to design systematic methods of data-preprocessing. Easy and efficient data preparation models may be designed which will significantly enhance capabilities of data mining algorithms.[3]

REFERENCES

[1].Class-Driven Attribute Extraction
Benjamin Van Durme, Ting Qian and Lenhart SchubertDepartment of Computer Science University of RochesterRochester, NY 14627, USA

[2]. A Self-Supervised Approach for Extraction of Attribute-Value Pairs from Wikipedia Articles Wladmir C. Brand˜ao1, Edleno S. Moura2, Altigran S. Silva2, and Nivio Ziviani1 1 Dep. of Computer

Science, Federal Univ. of Minas Gerais, BeloHorizonte, Brazil http://www.dcc.ufmg.br{wladmir,nivio}@dcc.ufmg.br2Dep. of Computer Science, Federal Univ. of Amazonas, Manaus, Brazil http://www.dcc.ufam.edu.br{edleno,alti}@dcc.ufam.edu.br

[3]. Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions Katharina Probst, RayidGhani, Marko Krema, Andrew Fano Accenture Technology Labs, Chicago, IL, USA Yan Liu Carnegie Mellon University, Pittsburgh, PA, USASHICHAO Z HANG and CHENGQI Z HANG "DATA PREPARATION FOR DATA MINING" Applied Artificial Intelligence, 17:375–381, 2003 Copyright # 2003 Taylor & Francis0883-9514/03 $12.00 +.00 DOI: 10.1080/08839510390219264

[4].Text Mining for Product Attribute ExtractionRayidGhani, Katharina Probst, Yan Liu1, Marko Krema, Andrew FanoAccenture Technology Labs 1Language Technologies Institute 161 N. Clark St, Chicago, IL Carnegie Mellon University, Pittsburgh,PA {rayid.ghani,katharina.a.probst,marko.krema,andrew.e.fano}@accenture.com,yanliu@cs.cmu.eduFRAME WORK FOR THE FIELD OF DATA MINING AND KNOWLEDGE DISCOVERY"International Journal of Information Technology & Decision Making