Exploring Classification & Clustering Techniques for Predictive Analytics

Manisha, Pawan Lathwal

Department of Computer Science, Department of Computer Science, B.M.I.E.T, Sonipat (Haryana)

ABSRACT: Diseases have been increased tremendously. Eight person from ten have diabeties, so the data in the hospital is increasing day by day. A huge amount of data has to be handled and for that we are using classification and clustering techniques which classifies the data and forms the clusters as well as we developed a tool for Apriori which predicts the strong rules. The tool is developed using .NET framework . For Clustering and Classification WEKA tool is used for extracting the results.

Keywords: Classification, Clustering, Predictive Analytics.

I. INTRODUCTION

Data Mining is the extraction of the knowledge from large amount of data.

Data Mining is a detailed process of analysing large amounts of data and picking out the relevant information. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. There are many types of Data Mining like Clustering, Classification, and Association Rules Mining.

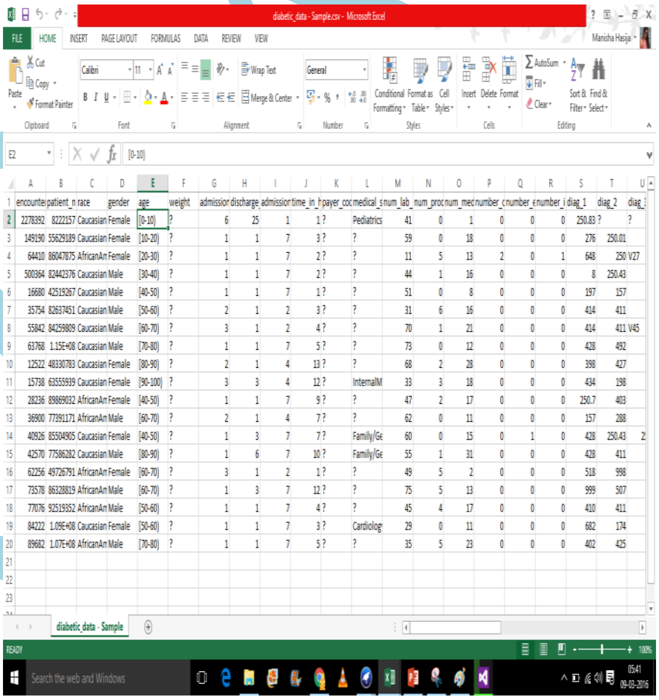
Classification in the data mining is to classify the data into similar and dissimilar classes. Clustering is the technique which is used for finding the clusters. Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, "if an item set is not frequent, any of its superset is never frequent".

II. CLASSIFICATION

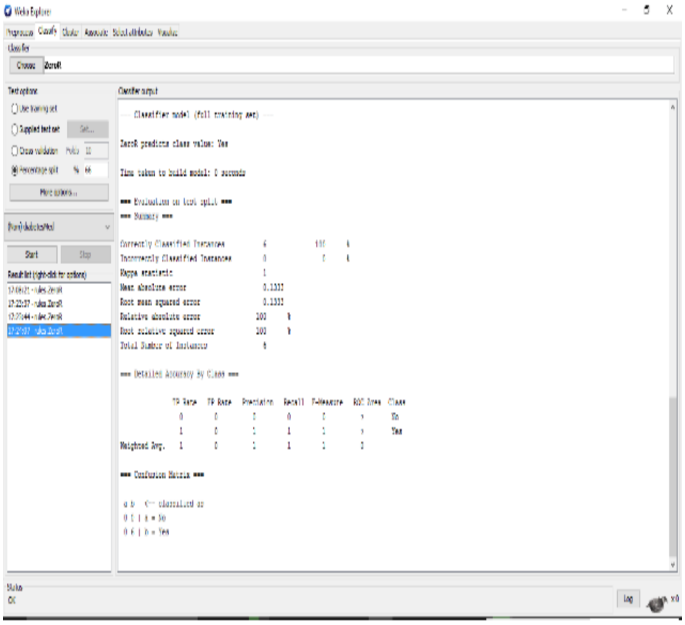
Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Here we use the diabetic data of the hospital in CSV format. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

ZeroR predicts the mean (for numeric class) or mode (for nominal class). Here ZeroR predicts class value and to build a model WEKA took 0 seconds.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	encounte	patient_n	race	gender	age	weight	admission	discharge_	admission	time_in_h	payer_coc	medical_s	num_lab_	num_proc	num_med	number_c	number_e	number_i	diag_1	diag_2	diag_3
2	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	?	Pediatrics	41	0	1	0	0	0	250.83	?	?
3	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	?	?	59	0	18	0	0	0	276	250.01	
4	64410	86047875	AfricanAm	Female	[20-30)	?	1	1	7	2	?	?	11	5	13	2	0	1	648	250	V27
5	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	?	?	44	1	16	0	0	0	8	250.43	
6	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	?	?	51	0	8	0	0	0	197	157	
7	35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3	?	?	31	6	16	0	0	0	414	411	
8	55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4	?	?	70	1	21	0	0	0	414	411	V45
9	63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5	?	?	73	0	12	0	0	0	428	492	
10	12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	?	?	68	2	28	0	0	0	398	427	
11	15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalM	33	3	18	0	0	0	434	198	
12	28236	89869032	AfricanAm	Female	[40-50)	?	1	1	7	9	?	?	47	2	17	0	0	0	250.7	403	
13	36900	77391171	AfricanAm	Male	[60-70)	?	2	1	4	7	?	?	62	0	11	0	0	0	157	288	
14	40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	?	Family/Ge	60	0	15	0	1	0	428	250.43	2
15	42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	?	Family/Ge	55	1	31	0	0	0	428	411	
16	62256	49726791	AfricanAm	Female	[60-70)	?	3	1	2	1	?	?	49	5	2	0	0	0	518	998	
17	73578	86328819	AfricanAm	Male	[60-70)	?	1	3	7	12	?	?	75	5	13	0	0	0	999	507	
18	77076	92519352	AfricanAm	Male	[50-60)	?	1	1	7	4	?	?	45	4	17	0	0	0	410	411	
19	84222	1.09E+08	Caucasian	Female	[50-60)	?	1	1	7	3	?	Cardiolog	29	0	11	0	0	0	682	174	
20	89682	1.07E+08	AfricanAm	Male	[70-80)	?	1	1	7	5	?	?	35	5	23	0	0	0	402	425	



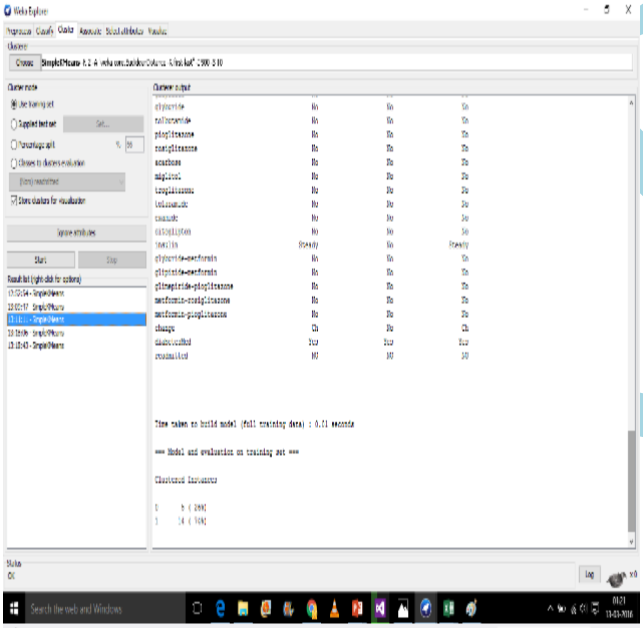
```
=== Classifier model (full training set) ===

ZeroR predicts class value: Yes

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances           6              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.1333
Root mean squared error                  0.1333
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0          0        0           ?       No
                 1         0          1          1        1           ?       Yes
Weighted Avg.    1         0          1          1        1           ?

=== Confusion Matrix ===

 a b   <-- classified as
 0 0 | a = No
 0 6 | b = Yes
```

a) ZeroR: It is the simplest classification method which relies on the target and ignores all predictors.
b) This classifier simply predicts the majority category (Class).
c) TP Rate: Rate of true positives (instances correctly classified as a given class).
d) FP Rate: Rate of false positives (instances falsely classified as a given class).

The Confusion Matrix for classification.

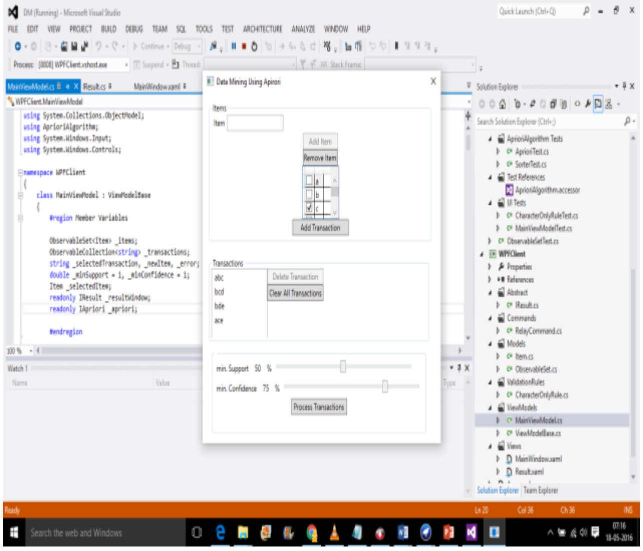
III. CLUSTERING

Clustering technique find groups of items that are similar. It divides a data set so that records with similar content are in the same group, and groups are as different as possible from each other. Cluster data using the k means algorithm. We can use either the Euclidean distance (default) or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean.



SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

```
glyburide                    No        No        No
tolbutamide                  No        No        No
pioglitazone                 No        No        No
rosiglitazone                No        No        No
acarbose                     No        No        No
miglitol                     No        No        No
troglitazone                 No        No        No
tolazamide                   No        No        No
examide                      No        No        No
citoglipton                  No        No        No
insulin                      Steady    No        Steady
glyburide-metformin          No        No        No
glipizide-metformin          No        No        No
glimepiride-pioglitazone     No        No        No
metformin-rosiglitazone      No        No        No
metformin-pioglitazone       No        No        No
change                       Ch        No        Ch
diabetesMed                  Yes       Yes       Yes
readmitted                   NO        NO        NO




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 26%)
1      14 ( 74%)
```



```
using System.Collections.ObjectModel;
using AprioriAlgorithm;
using System.Windows.Input;
using System.Windows.Controls;

namespace WPFClient
{
    class MainViewModel : ViewModelBase
    {
        #region Member Variables

        ObservableSet<Item> _items;
        ObservableCollection<string> _transactions;
        string _selectedTransaction, _newItem, _error;
        double _minSupport = 1, _minConfidence = 1;
        Item _selectedItem;
        readonly IResult _resultWindow;
        readonly IApriori _apriori;

        #endregion
```

Items: ☐ a, ☐ b, ☑ c

Transactions: abc, bcd, bde, ace

min. Support 50 %

min. Confidence 75 %

This model took 0.01 second to build.Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and “gender").

IV. APRIORI TOOL

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

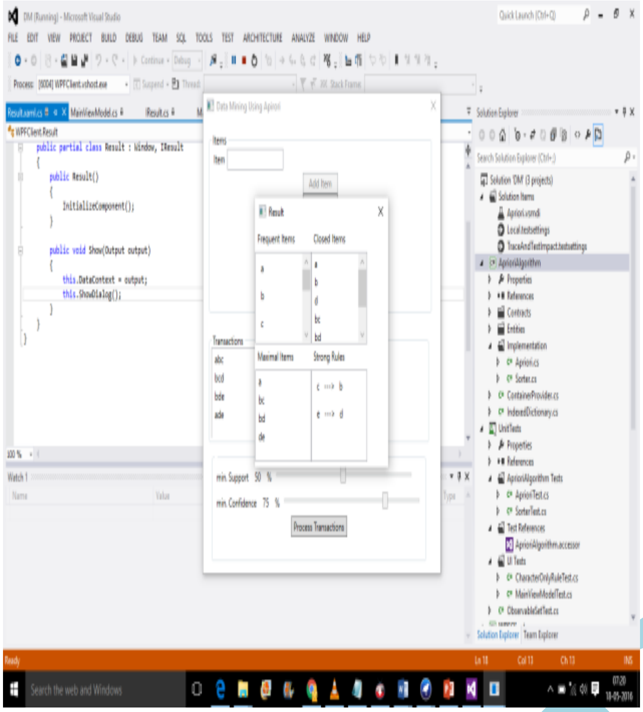
```
$F_l$ = (Frequent Itemsets);
for ($k$=1; $F_k \neq Ø$; k++) do begin
    $C_{k+1}$ =Apriori_gen ($F_k$); // New Candidates
for all transactions $t \in$ Database do begin
        $C'_t$ = subset ($C_{k+1}$, t); // Candidates Contained in $t$
for all candidates c $\in C'_t$ do
c.count++;
end
$F_{k+1}$ ={C $\in C_{k+1}$c.count>= minimum support}
end
end
Answer $U_kF_k$;
```

a	b
0	0
0	6

The tool was developed on visual studio. We are using Visual Studio .NET tool, and the coding is done in C Sharp (C#). For Front End WPS (Windows Presentation Foundation) is used. WPS is a powerful framework for building Windows applications.WPF employs XAML, an XML-based language, to define and link various interface elements.

The main window has options to add and delete transactions as well as we can set minimum support and minimum confidence on this page. If process transaction is clicked after selecting transaction, the result window will appear.

The result window will calculate strong rules by frequent items. It will find all frequent itemsets and get frequent items,Items whose occurrence in database is greater than or equal to the min.support threshold. Get frequent itemsets:Generate candidates from frequent items. Then prune the results to find the frequent itemsets and will generate strong association rules from frequent itemsets, rules which satisfy the min.support and min.confidence threshold.



```
public partial class Result : Window, IResult
{
    public Result()
    {
        InitializeComponent();
    }

    public void Show(Output output)
    {
        this.DataContext = output;
        this.ShowDialog();
    }
}
```

Data Mining Using Apriori

Items

Item

Add Item

Transactions

abc
bcd
bde
ade

min. Support 50 %

min. Confidence 75 %

Process Transactions

Result

Frequent Items	Closed Items
a	a
b	b
c	d
	bc
	bd

Maximal Items	Strong Rules
a	c ---> b
bc	e ---> d
bd	
de	

We took min sup 50% and min confi 75% after taking transactions as abc, bcd, bde, ace and from this strong rules are c→b, e→d.

V. CONCLUSION

The paper shows the clustering & classification rule techniques for predictive analysis and evaluates the performance of association rule mining algorithms in the context of database partitioning. The focuses on Apriori and sampling algorithms for frequent item sets mining when the data is partitioned into a number of given segments. Apriori algorithm scans the database multiple number of times for counting the support for the item sets. The strong rules can be predicted by using this tool. We shows the clustering & classification of data on WEKA and it classified the data perfectly & shows the confusion matrix.

VI. REFERENCES

[1]. Association Rule Learning – Wikipedia, the free encyclopedia.

[2]. Chen, H., Zhan, Y., Li, Y., (2010), "The application of Decision Tree in Chinese email Classification". In the proceeding of 9th International Conference on machine Learning and Cybernetics, 2010, pp. 305-308.

[3]. Feng Yucai, "Association Rules Incremental Updating Algorithm", Journal of Software, Sept., 1998

[4]. Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers.

[5]. Execution of APRIORIAlgorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[6]. Sonam Narwal , Comparison the Various Clustering and Classification Algorithms of WEKA Tools, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.

[7]. [7] Nimrat Kaur Sidhu, Rajneet Kaur "Clustering In Data Mining "International Journal of Computer Trends and Technology (IJCTT),V4(4):710-714 April Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.

[8]. Huang, A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

[9]. [9] Shraddha, A Literature Review on Data Mining and its Techniques, Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

[10]. Antonie, Application of Data Mining Techniques for Medical Image Classification, SIGKDD conference, 2001

[11]. Baluni, A comparative study of various approaches to explore factors for vehicle collision, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 3, September – October 2012.

[12]. R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules," In Proc. of VLDB '94, pp. 487-499, Santiago, Chile, Sept. 1994.