IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. I (Jan.-Feb. 2017), PP 11-18
www.iosrjournals.org

Different Classification Technique for Data mining in Insurance Industry using weka

Mohd Ahtesham Farooqui[1], Dr. Jitendra Sheetlani[2]
[1](Research scholar) Department of computer science & application Sri Satya Sai University of Technology & Medical Sciences
[2](Associate Professor) Department of computer science & application Sri Satya Sai University of Technology & Medical Sciences

Abstract: *this paper addresses the issues and techniques for Property/Casualty actuaries applying data mining methods. Data mining means the effective unknown pattern discovery from a large amount database. It is an interactive knowledge discovery procedure which is includes data acquisition, data integration, data exploration, model building, and model validation. The paper provides an overview of the data discovery method and introduces some important data mining method for application to insurance concluding cluster discovery approaches.*
Keywords: *kdd,weka,gnu*

I. Introduction

Because of the rapid progress of information technology, the amount of information stored in insurance databases is rapidly increasing. These large databases include data and constitute wealth of a potential valuable business information goldmine. As novel and developing loss exposures emerge in the ever-changing insurance atmosphere and insurance databases change structure. In addition, new applications such as dynamic financial analysis and catastrophe modeling require the storage, retrieval, and analysis of complex multimedia objects. Finding the valuable information hidden in those databases and identifying appropriate models is a difficult task. Analysis of cluster is a common method that is often using in using huge data sets. Originating from statistics area, most algorithms for analysis of cluster have originally been developed for relatively less data sets. In recent years, the clustering algorithms have been extended to efficiently work on large data sets, and some of them even allow the clustering of high-dimensional feature vectors. Analysis of Decision tree is another famous data mining method that can be used in numerous actuarial practice areas. We discuss how to use decision trees to make important design decisions and explain the interdependencies among the properties of insurance data. We will also provide examples of how data mining techniques can be used to improve the effectiveness and efficiency of the modeling process.

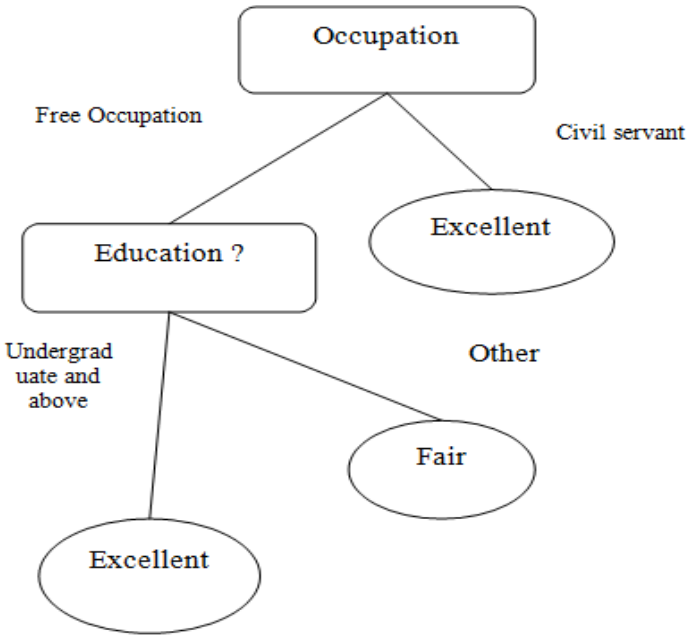


Figure 1. Decision tree example of Insurance customer Systems

II. Data mining

Knowledge discovery and Data mining from data bases has received additional attention in the present years. Data mining, the hidden predictive data mining from huge amount of databases, is a powerful novel technology with high potential to help company's attention on most important knowledge in their data warehouses. KDD is the non-trivial identifying valid process, earlier unknown and potentially valuable patterns in the data.

These patterns are used to generate classifications or predictions new data, describe existing data, summarize huge database contents to the support decision creating and provide graphical data visualization Discovering valuable patterns hidden in a database achieve an essential role in many data frequent pattern mining, and high utility pattern mining [1].

Here, itemset meaning utility is interestingness, item profitability or importance to users. Items Utility in a transaction database two aspects consists:

1. The importance of distinct items, which is known as external utility, and
2. The importance of items in transactions, which is known as internal utility.

Table 1.1: An Example Database

TID	Transaction TU
T1	(A,1) (C,10) (D,1) 17
T2	(A,2) (C,6) (E,2) (G,5) 27
T3	(A,2) (B,2) (D,6) (E,2) (F,1) 37
T4	(B,4) (C,13) (D,3) (E,1) 30
T5	(B,2) (C,4) (E,1) (G,2) 13
T6 (	A,1) (B,1) (C,1) (D,1) (H,2) 12

III. Data mining process

Data mining combines method from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, interrelations concept, and interesting patterns automatically from huge corporate databases. Its basic aim is to mining data from information to support the decision-making procedure. Two primary data mining functions are:

prediction, which conclude finding unknown values/relationships/patterns from known values; and description, which provides large database interpretation. A data mining procedure commonly conclude the following four various steps.

STEP 1: Data acquisition. Although a final data set has been generated for discovery in few applications, DM can be achieved on a variables subset or data samples in an advance database.

STEP 2: Preprocessing data. Once the final data is selected, data is then preprocessed for cleaning, scrubbing, and transforming to increase the discovery effectiveness. At the time of this preprocessing step, developers eliminate the noise or outliers if necessary and decide on approaches for dealing with missing information fields and accounting for time sequence information or known changes. In addition, the data is often transformed to decrease the efficient variables number under consideration through either converting one data type.

STEP 3: Data exploration and model building. The third DM step refers to activities series for example deciding on the DM operation type; selecting the DM method; choosing

IV. Classification Technique

Decision tree induction:

A decision tree is a flowchart, for example, tree structures, in which every single internal nodes refer a test on a trait, every branch speaks to a test result, what not leaf node holds a class label. [6].

Advantages: Amongst numerous data mining method, decision trees have many advantages [7]. Decision trees are easily to interpret and understand. They need some data and are able to handle both categorical and numerical data. They are robust in nature, therefore, they achieve well even if its assumptions are somewhat violated through true model from which information were created. Decision trees archive well with huge data in a short time. Huge data amounts can be analyzed applying personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

Limitations: The learning issue an optimal decision tree is called as NP-complete. Such technique can't ensure to return overall optimal decision tree. Decision tree learners make over-complex trees that don't information sum up well. Components for instance pruning are important to evade this issue.

Nearest neighbor classifier

The k-NN is technique for ordering objects in view of nearest preparing tests in highlight space. K-NN is an example based learning strategy, or lethargic learning. It can likewise be relapse reason utilized. The k-closest neighbor calculation is among each machine-learning calculation. The space is divided into districts by specific areas and preparing tests marks. A point in space is doled out to class c on the off chance that it is the most

incessant class name among k closest preparing. Traditionally Euclidean separation is utilized as separation metric; however this will just work with numerical qualities.

Artificial neural network

NN are analytic method modeled after hypothesized learning processes in the cognitive framework and the neurological brain works and predicting capable novel observations (on specific variables) from different perceptions then executing a taking in strategy from existing information. NN is one of the Data Mining strategies. The primary level is to plan specific system engineering. System is then subjected to the "preparation" strategy. In that level, neurons using a procedure to the numerous inputs to the network weights adjust in order to optimally predict data on which "training" is performed. The final outcomes of "network" established in "learning" process represents a pattern detected in the particular information.

Support vector machines

SVM were first brought to solve pattern type and regression issue through Vapnik and his colleagues [8] [6]. Viewing enters facts as vectors units in an n-dimensional area, a SVMs will construct a isolating hyper-aircraft in that area, one which maximizes margin among the 2 different facts units [6]. to calculate margin, various parallel hyper-planes are constructed, one on all aspect of the keeping apart hyper-plane, that are "driven up against" the two records units [6]. an awesome separation is carried out via hyper-aircraft that has the largest distance to the neighboring information factors of each lessons, on account that in commonplace large the margin [6]. this hyperplane is discovered thru applying the support -vectors and margins.

Genetic Programming

GP has been utilized as a part of the exploration field recent years to determine information mining issue. The reason GP is so extensively used is fact that prediction rules are most naturally represented in GP. Also, GP has make great results with worldwide environment. The search space for classification can be defined as containing various ‘peaks’, this causes local search algorithms, for example simulated annealing, to achieve badly.

GP comprises of stochastic search algorithms in view of reflections of the Darwinian procedures advancement. All candidate solution arrangement through individual in GP. The arrangement is coded into chromosome for instance structures that can be changed and additionally consolidated with couple of other individual's chromosome. Every individual incorporate a fitness value, which measures singular quality, at the end of the day how close candidate solution is from being optimal. Based on fitness value, individuals are selected to mate. This procedure generates a novel individual through combining two or an additional chromosome, this procedure is known as crossover.

They are joined with all other with the expectation that these novel individuals will develop and get to be higher than parents. There are various parameters used to decide when the calculation ought to stop, and all information set can have extremely different settings.In each case, the best individual is stored across generations and is returned when the algorithm stops. The most usually used parameter is various generations. [9]

V. Literature survey

K. Umamaheswari (2014) et al presents that in global era, Insurance systems various tremendous developments in society. Because of enhanced stress in day-to-day life, the insurance increased demand growth. The purpose of the paper goals to present how data mining is valuable in insurance industry, how its methods create better outcomes in insurance sector and how data mining improve in decision making with insurance data. The conceptual paper is written based on secondary study, observation from various journals, magazines and reports [2].

A. B. Devale (2012) et al present that information discovery in financial organization have been constructed and operated generally to support decision making applying knowledge as strategic factor. In this paper, we investigate use of numerous data mining methods for data discovery in insurance business. Existing software are inefficient in present such data characteristics. Proposed information mining strategies, the decision- maker can describe the insurance activities expansion to enable the different powers in existing life coverage division [3].

Ruxandra PETRE (2013) et al present that Over the past years, data mining became a matter of considerable importance because of huge data amounts presented in the applications belonging to numerous domains. Data mining, a dynamic and fast-expanding field, that applies advanced data analysis methods. In order to the discover relevant patterns, relations and trends contained within data, knowledge impossible to observe applying other method. The paper focuses on presenting the data mining applications in the business atmosphere. It include a common data mining overview, providing a concept definition, enumerating six

primary data mining method and mentioning basic fields for which data mining can be using. The paper also presents main business areas which can advantage from data mining tools use, along with their use cases: retail, insurance and banking. Also the basic commercially available data mining tools and their key features are presented within paper. Besides data mining analysis and business areas that can effectively using it, the paper presents basic features of a data mining solution that can be using for business atmosphere and the architecture, with its main components, for solution, that would help improve customer experiences and decision-making [4].

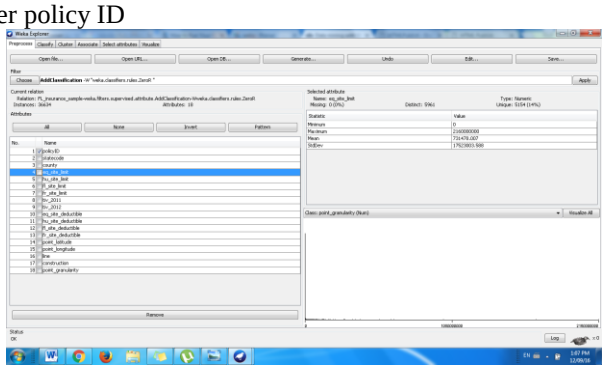
Weka tool

Weka is an acclaimed machine learning programming suite in Java form written, generated at the Waikato University, New Zealand. Weka is free programming present under GNU basic Public License. The Weka workbench includes tools for visualization and algorithms collection for analysis and predictive modeling of data purpose, together with the graphical UIs for easily access to this functionality. Weka is a machine learning algorithms set for explaining true information mining issue. It is composed in Java and keeps running on any platform. Weka underpins different standard information mining tasks, extra particularly, data preprocessing, and clustering, characterization, regression, visualization, and feature selection. Wekas provide access to SQL databases using Java Database Connectivity and can technique outcomes returned with applying database question [5].

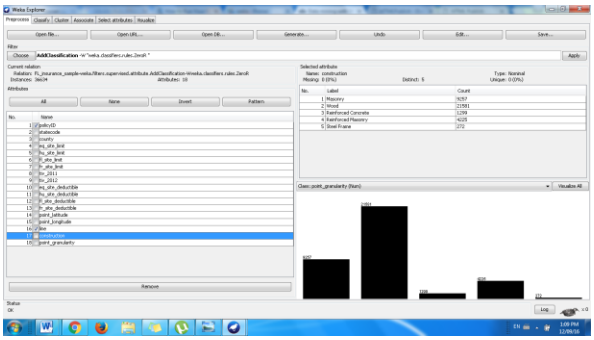
VI. Result and simulation

Classifier

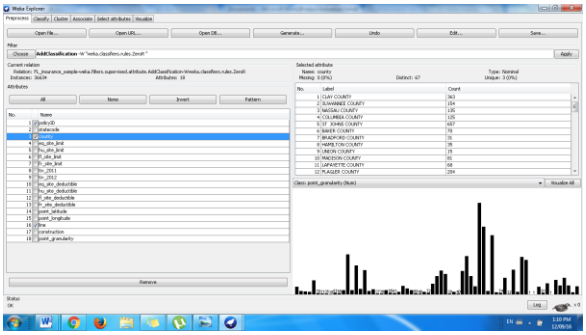
Addclassification ZeroR over policy ID



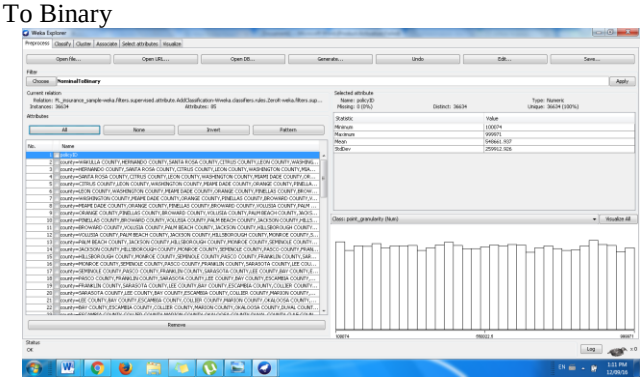
Policy id and line

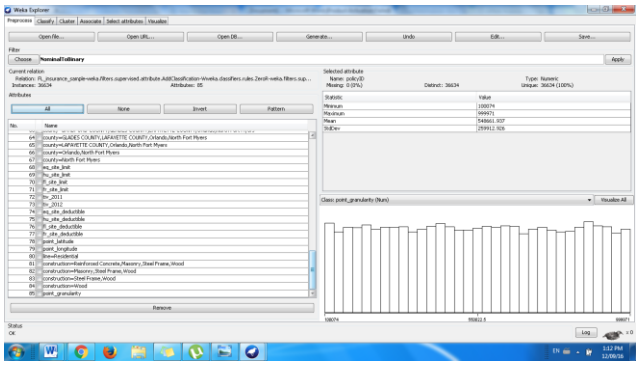


Policy ID country and line

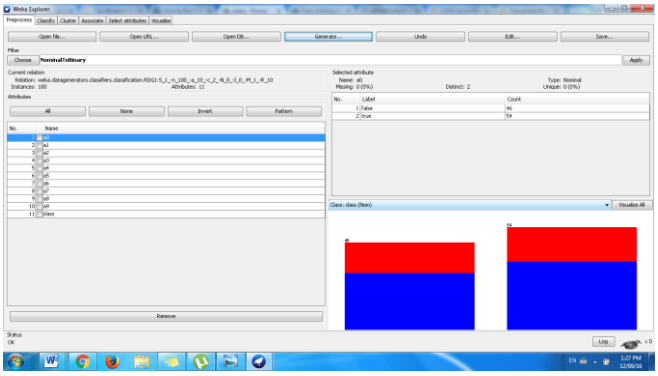


When we apply Nominal To Binary





Nominalbinary



Nomil To Binary
True: 46
False: 54

Classify j-48 tree
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances	28	82.3529 %
Incorrectly Classified Instances	6	17.6471 %
Kappa statistic	0.6357	
Mean absolute error	0.1909	
Root mean squared error	0.3895	
Relative absolute error	40.9445 %	
Root relative squared error	77.4685 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.85	0.214	0.85	0.85	0.85	0.859	c0
	0.786	0.15	0.786	0.786	0.786	0.859	c1
Weighted Avg.	0.824	0.188	0.824	0.824	0.824	0.859	

=== Confusion Matrix ===

```
 a  b   <-- classified as
17  3 |  a = c0
 3 11 |  b = c1
```

Tree formation
J48 pruned tree

```
a5 = false: c0 (44.0/2.0)
a5 = true
|   a8 = false
|   |   a9 = false
|   |   |   a2 = false: c0 (4.0/1.0)
|   |   |   a2 = true
|   |   |   |   a0 = false
|   |   |   |   |   a4 = false: c1 (2.0)
|   |   |   |   |   a4 = true: c0 (2.0)
|   |   |   |   a0 = true: c1 (5.0)
|   |   a9 = true: c1 (15.0/1.0)
|   a8 = true
|   |   a1 = false
|   |   |   a2 = false
|   |   |   |   a0 = false: c1 (4.0/1.0)
|   |   |   |   a0 = true: c0 (3.0)
|   |   |   a2 = true
|   |   |   |   a4 = false: c1 (5.0)
|   |   |   |   a4 = true: c0 (2.0)
|   |   a1 = true: c0 (14.0/2.0)
```

Number of Leaves : 11

Size of the tree : 21

BFTREE:

Time taken to build model: 0.12 seconds

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

Correctly Classified Instances	28	82.3529 %
Incorrectly Classified Instances	6	17.6471 %
Kappa statistic	0.6277	
Mean absolute error	0.1791	
Root mean squared error	0.3964	
Relative absolute error	38.4213 %	
Root relative squared error	78.8413 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.9	0.286	0.818	0.9	0.857	0.852	c0
	0.714	0.1	0.833	0.714	0.769	0.852	c1
Weighted Avg.	0.824	0.209	0.824	0.824	0.821	0.852	

=== Confusion Matrix ===

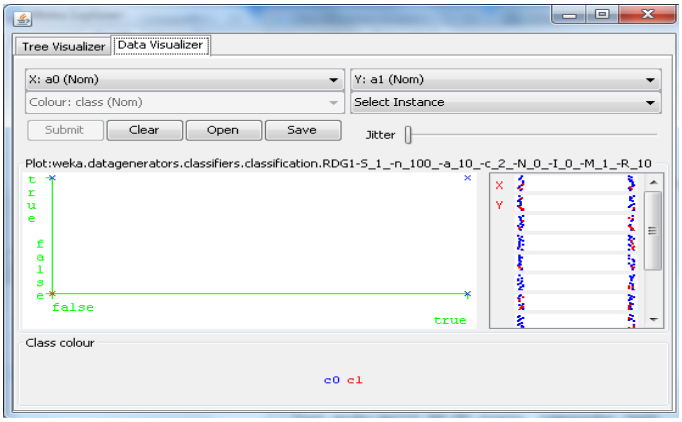
```
  a  b   <-- classified as
 18  2 |  a = c0
  4 10 |  b = c1
```

Tree
Best-First Decision Tree

```
a5=(true)
| a8=(false)
| | a9=(true)
| | | a3=(true): c1(9.0/0.0)
| | | a3!=(true)
| | | | a0=(false): c1(3.0/0.0)
| | | | a0!=(false): c1(2.0/1.0)
| | a9!=(true)
| | | a2=(true)
| | | | a0=(true): c1(5.0/0.0)
| | | | a0!=(true)
| | | | | a4=(false): c1(2.0/0.0)
| | | | | a4!=(false): c0(2.0/0.0)
| | | a2!=(true)
| | | | a3=(true): c0(1.0/1.0)
| | | | a3!=(true): c0(2.0/0.0)
| a8!=(false)
| | a1=(false)
| | | a2=(true)
| | | | a4=(false): c1(5.0/0.0)
| | | | a4!=(false): c0(2.0/0.0)
| | | a2!=(true)
| | | | a0=(false): c1(3.0/1.0)
| | | | a0!=(false): c0(3.0/0.0)
| | a1!=(false)
| | | a7=(true)
| | | | a0=(false): c1(2.0/1.0)
| | | | a0!=(false): c0(3.0/0.0)
| | | a7!=(true): c0(8.0/0.0)
a5!=(true)
| a1=(true): c0(18.0/2.0)
| a1!=(true): c0(24.0/0.0)
```

Size of the Tree: 33

Number of Leaf Nodes: 17
User classisifer



Time taken to build model: 18.24 seconds

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

Correctly Classified Instances	20	58.8235 %
Incorrectly Classified Instances	14	41.1765 %
Kappa statistic	0	
Mean absolute error	0.4652	
Root mean squared error	0.504	
Relative absolute error	99.7807 %	
Root relative squared error	100.2422 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.588	1	0.741	0.5	c0
	0	0	0	0	0	0.5	c1
Weighted Avg.	0.588	0.588	0.346	0.588	0.436	0.5	

=== Confusion Matrix ===

a	b	<-- classified as
20	0	a = c0
14	0	b = c1

VII. Conclusion

In this paper we study about data mining techniques over insurance data after that insurance data classified using different classifier and came to now that bftree perform well.

References

[1]. Smita R. Londhe, Rupali A. Mahajan and Bhagyashree J. Bhoyar," Overview on Methods for Mining High Utility Itemset from Transactional Database", International Journal of Scientific Engineering and Research (IJSER) www.ijser.in, Volume 1 Issue 4, December 2013

[2]. K. Umamaheswari and Dr. S. Janakiraman," Role of Data mining in Insurance Industry", An international journal of advanced computer technology, 3 (6), June-2014 (Volume-III, Issue-VI), pp: 961- 966.

[3]. A. B. Devale and Dr. R. V. Kulkarni," APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.4, July 2012, pp: 31- 40.

[4]. Ruxandra PETRE," Data Mining Solutions for the Business Environment", Database Systems Journal vol. IV, no. 4/2013, pp:21-31

[5]. Dr. Sudhir B. Jagtap and Dr. Kodge B. G," Census Data Mining and Data Analysis using WEKA", (ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore, Census Data Mining and Data Analysis using WEKA, pp: 35-40.

[6]. Jiawei Han, Michelin Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, USA, 2001, 70-181.

[7]. Lyman, Peter; Hal R. Varian (2003). "How Much Information"

[8]. "A training algorithm for optimal margin classifiers" B.E. Boser, I.M.Guyon and V.N.Vapnik, The 5th annual workshop on computational learning theory, 1992.

[9]. Genetic Programming, John R. Koza, MIT Press, 1998.