International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

Implementation and Evaluation of Rule Induction Algorithm with Association Rule Mining: A study in life insurance

Kapil Sharma, Sheveta Vashisht, Heena Sharma, Jasreena kaur Bains, Richa Dhiman

Computer Science and Engineering, Lovely Professional University, Punjab, India

kapilsharma701@gmail.com

Computer Science and Engineering, Lovely Professional University, Punjab, India

sheveta.16856@lpu.co.in

Computer Science and Engineering, L.L.R.I.E.T Moga(PTU), Punjab, India

heenasharma103@gmail.com

Computer Science and Engineering, Lovely Professional University, Punjab, India

jasreenakaur05@gmail.com

Computer Science and Engineering, Lovely Professional University, Punjab, India

richadhiman58@gmail.com

Abstract:

Data Mining: extracting useful insights from large and detailed collections of data. With the increased possibilities in modern society for companies and institutions to gather data cheaply and efficiently, this subject has become of increasing importance. This interest has inspired a rapidly maturing research field with developments both on a theoretical, as well as on a practical level with the availability of a range of commercial tools.

In this research work we use rule induction in data mining to obtain the accurate results with fast processing time. We using decision list induction algorithm to make order and unordered list of rules to coverage of maximum data from the data set. Using induction rule via association rule mining we can generate number of rules for training dataset to achieve accurate result with less error rate. We also use induction rule algorithms like confidence static and Shannon entropy to obtain the high rate of accurate results from the large dataset. This can also improves the traditional algorithms with good result.

Keywords: rule induction, association rule mining, decision list induction, Shannon entropy, data mining, confidence static

1 INTRODUCTION

Data mining techniques are the result of a long process of research and product development. This evolution began when Business data was first stored on computers, continued with improvements in data access and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

1.1 Separate and Conquer paradigm:

Among the rule induction methods, the "separate and conquer" approaches are very popular during the 90's. The goal is to learn a prediction rule from data If Premise Then Conclusion « Premise » is a set

of conditions « attribute – Relational Operator – Value ». For instance, Age > 45 and Profession = Workman

In the supervised learning framework, the attribute into the conclusion part is of course the target attribute. A rule is related to only one value of the target attribute. But one value of the target attribute may be concerned by several rules.

1.2 Compared to classification tree algorithms:

Which are based on the divide and conquer paradigm, their representation bias is more powerful because it is not constrained by the arbores cent structure. It needs sometimes a very complicated tree to get an equivalent of a simple rule based system. Some splitting sequences are replicated into the tree. It is known as the "replication problem".

1.3 Compared to the predictive association rule algorithms:

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

They do not suffer of the redundancy of the induced rules. The idea is even to produce the minimal set of rules which allows classifying accurately a new instance. It enables to handle the problem of collision about rules, when an instance activates two or several rules which lead to inconsistent conclusions.

We describe first two separate and conquer algorithms for the rule induction process. Then, we show the behavior of the classification rules algorithms implemented by a tool.

Separate and Conquer algorithms

- **Induction of ordered rules(Decision list induction)**
- **Induction of unordered rules**

Induction of ordered rules (Decision list induction)

The induction process is based on the top down separate and conquers approach. We have nested procedures that are intended to create the set of rules from the target attribute, the input variables and the instances.

The rule based system has the following structure:

IF Condition 1 Then Conclusion 1

Else If Condition 2 Then Conclusion 2

Else If…

Else If (Default rule) Conclusion M

Decision list induction algorithm:

Decision List (target, inputs, instances)

Ruleset = ∞

Repeat

Rule = Specialize (target, inputs, instances)

If (Rule != NULL) Then

Ruleset = Ruleset + {Rule}

Instances = Instances – {Instances covered by the rule}

End if

Until (Rule = NULL)

Ruleset = Ruleset + {Default rule (instances)}

Return (Ruleset)

Induction of unordered rules:

Ordered set of rules, when we read the i-th rule, we must consider the (i-1) preceding rules. It is impracticable when we have a large number of rules.

The classifier is now outlined as the following:

If Condition 1 Then Conclusion 1

If Condition 2 Then Conclusion 2

…

(Default rule) Conclusion M

(Ruleset)

2 PREVIOUS WORKS

There are number of practical works have been presented where most existing rule induction algorithms are used. Authors in [1] proposed Discovery of spatial association rules in georeferenced census data. It was relational mining approach. Authors in [3, 4] proposed Top down induction of model trees with regression and splitting nodes and Ranking Mechanisms in Metadata Information Systems for Geospatial Data. Authors in [8] proposed Rule Induction with CN2 with Some recent improvements over traditional algorithms. They also proposed post pruning and hybrid pruning technique along with rule induction method to obtain high rate of accurate results. They also reduced the induced set of rules and computational time with high coverage of data from large data set. They also used decision tree and rule induction method with the help of data mining software.

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

3 SOLUTIONS

3.1 Induction of Ordered Rules

Dataset

We take life insurance policy data; we want to detect the customers who having good policy based on customer categories and we have to obtain accurate result with less computational time.

Importing the database

After the launching of Tanagra, we create a new diagram by clicking on the FILE / NEW menu. We import the life insurance .xls file.

Sampling Algorithm

We want to subdivide the dataset into a learning sample (50%) and a test sample. We use the SAMPLING Component.

Sampling Algorithm

We want to subdivide the dataset into a learning sample (50%) and a test sample. We use the SAMPLING Component.

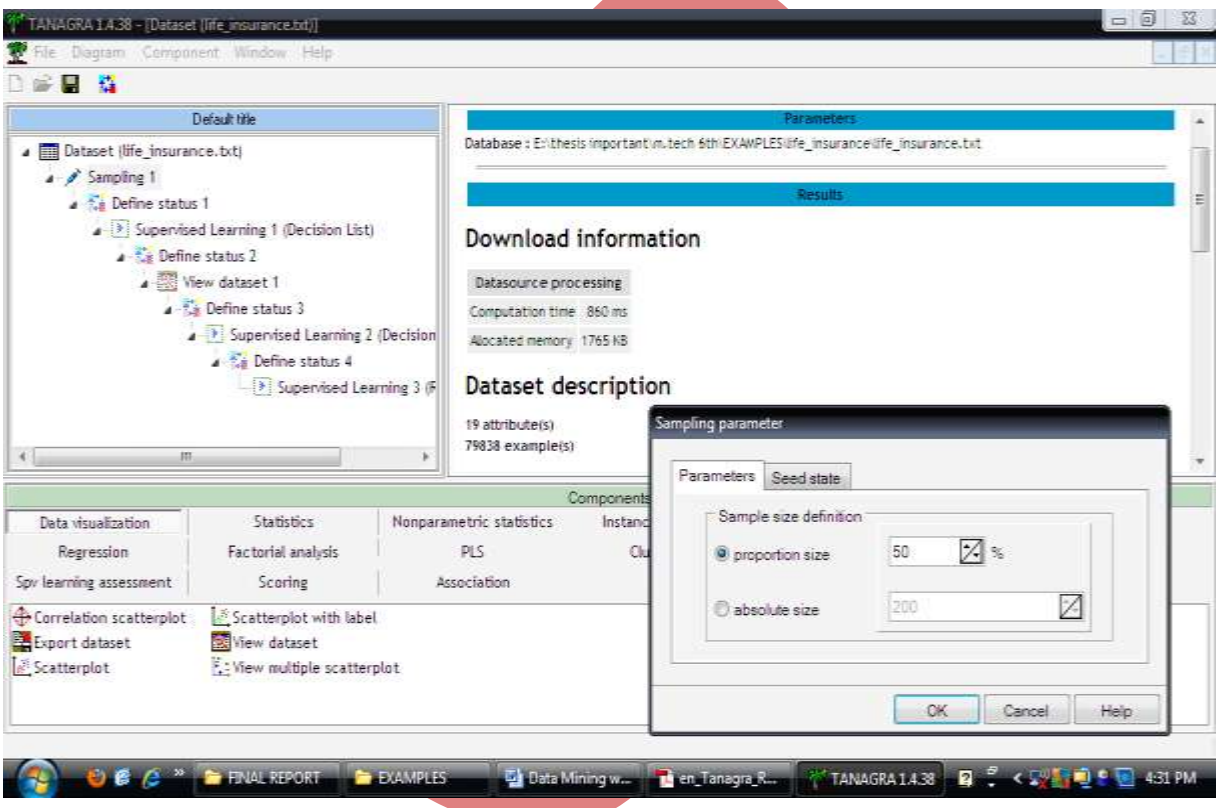


Fig 1

We set now “target” as TARGET attribute and the others as INPUT ones using the DEFINE STATUS component.

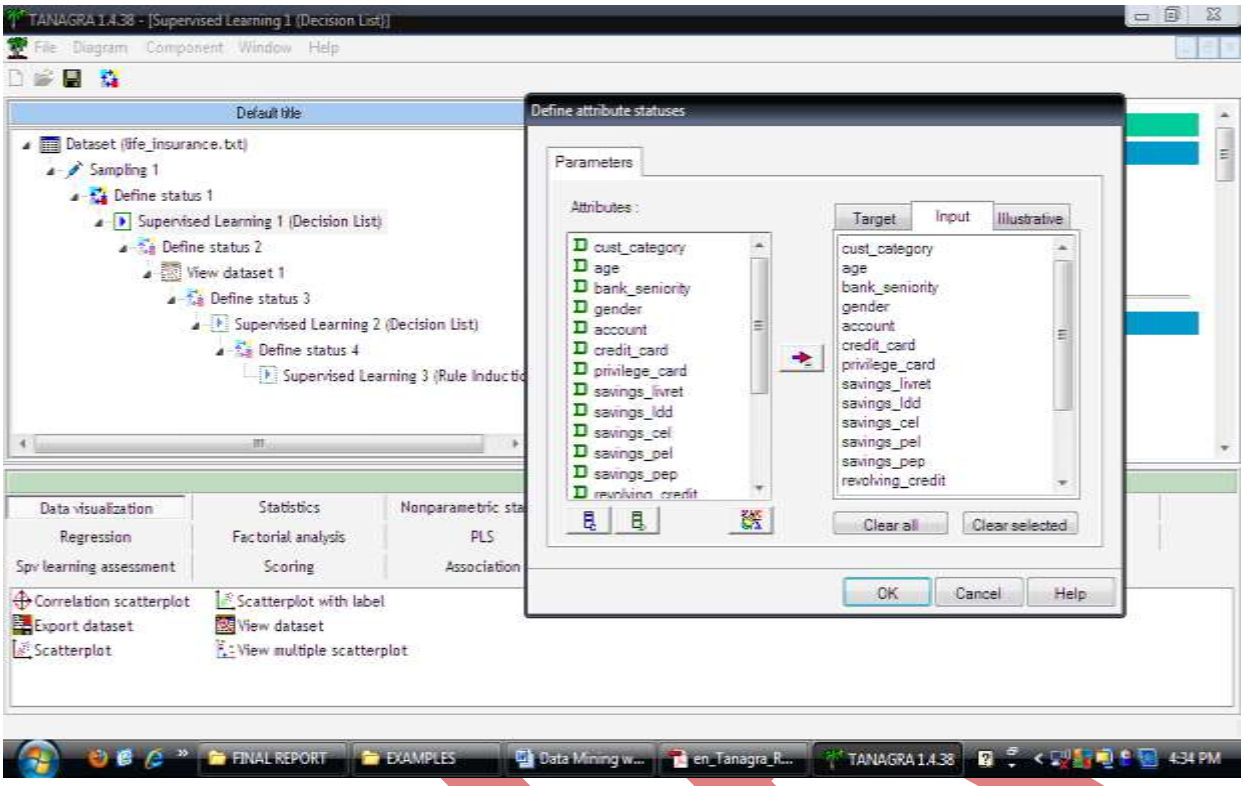


Fig 2

Induction of Decision Lists

We add the DECISION LIST component into the diagram. We click on the SUPERVISED PARAMETERS menu, **the J-MEASURE** is the default measure.

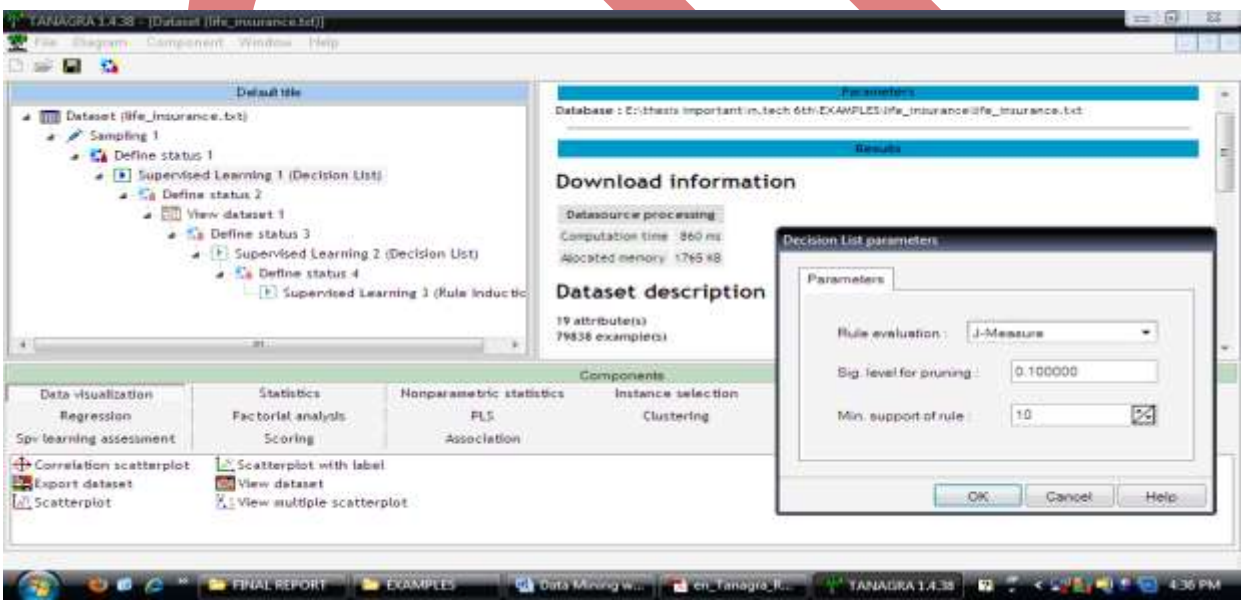


Fig 3

We validate these settings and we click on the VIEW menu. We obtain **20 rules in 703 ms.**

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

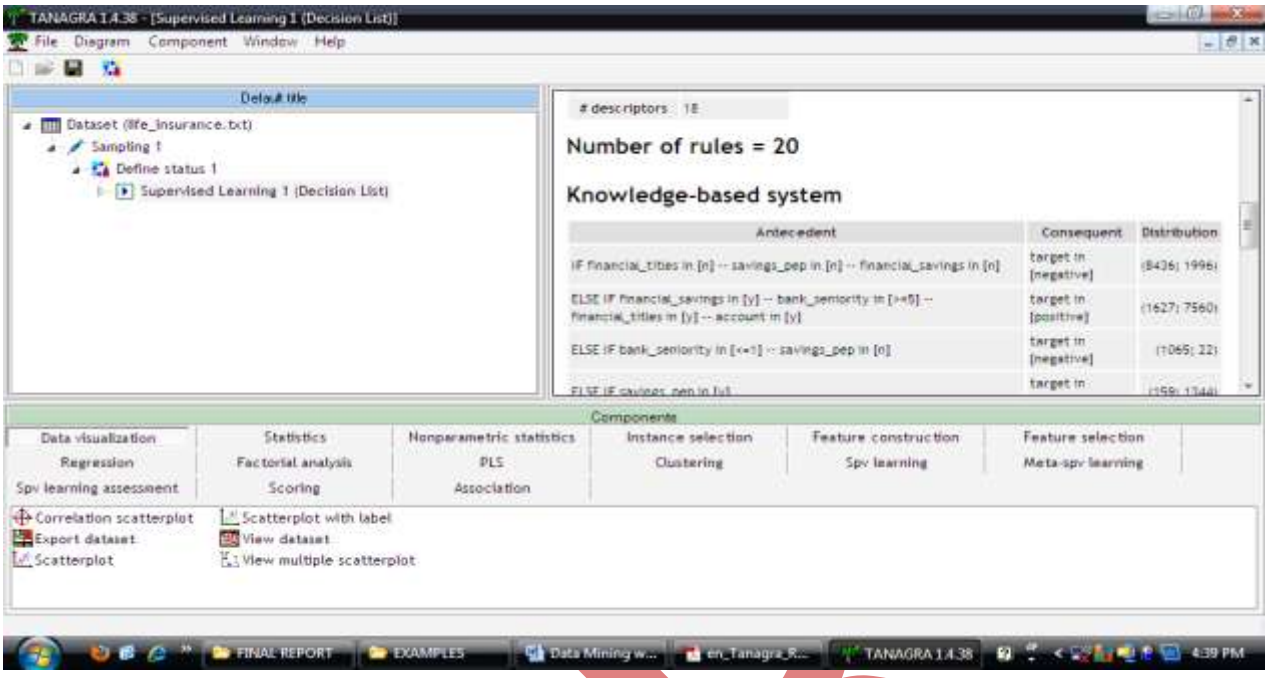


Fig 4

Modifying the parameters of the learning algorithm

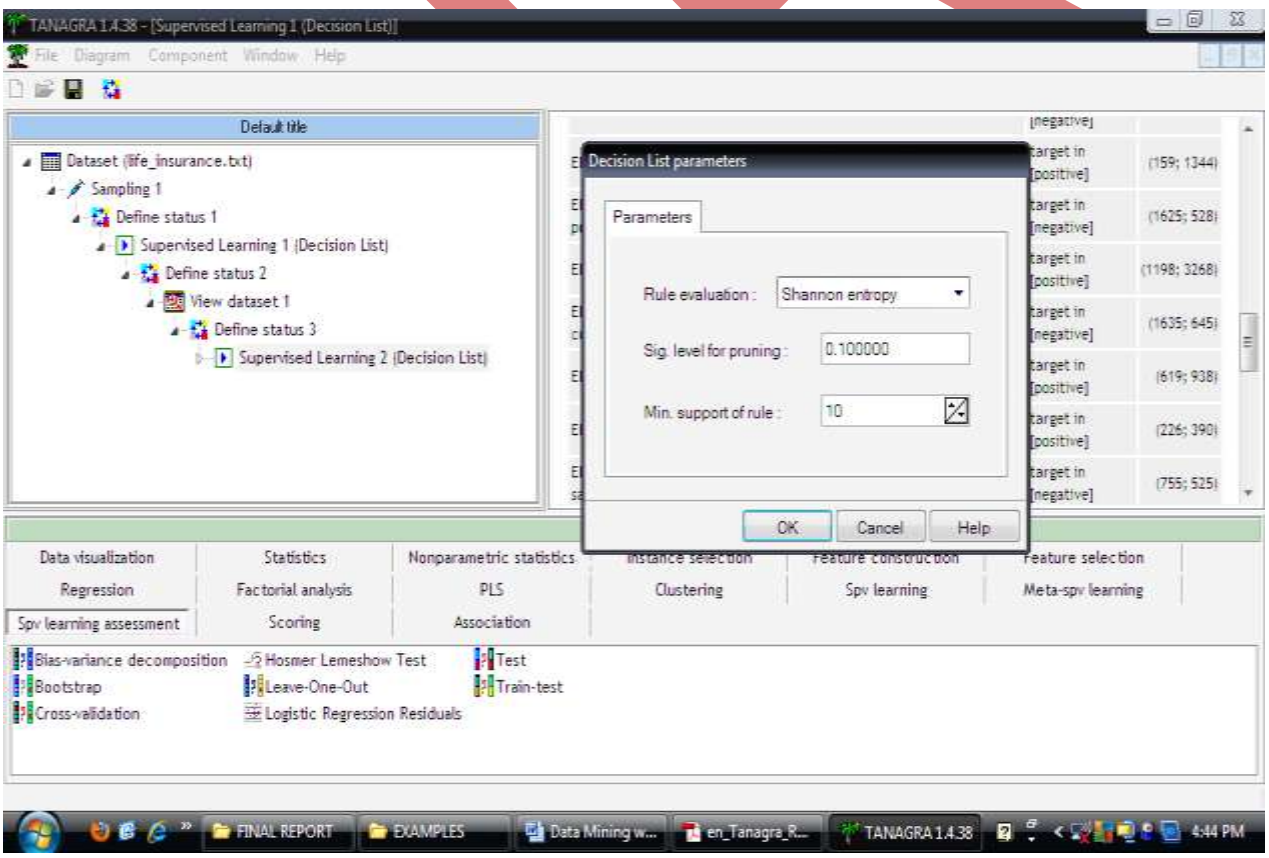


Fig 5

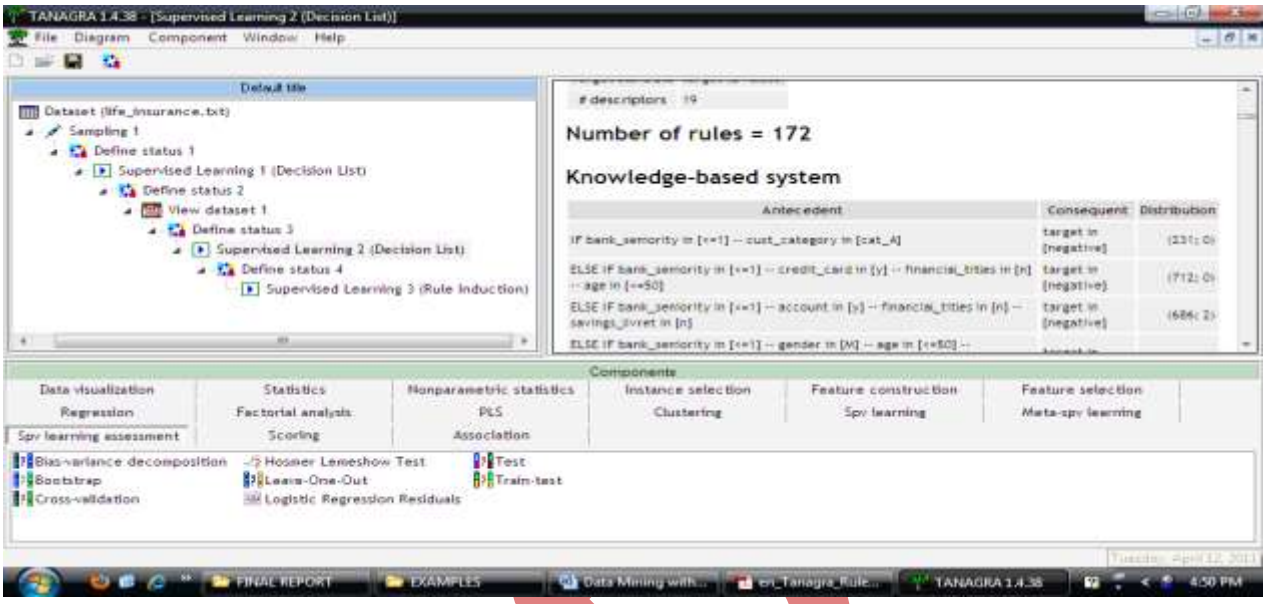


Fig 6

We validate these settings and we click on the VIEW menu. We obtain **172 rules in 9203 ms**.

3.2 Induction of unordered rules

We use the RULE INDUCTION component (SPV LEARNING tab) in order to generate a set of unordered rules. We click on the SUPERVISED PARAMETERS menu, the default settings are the following.

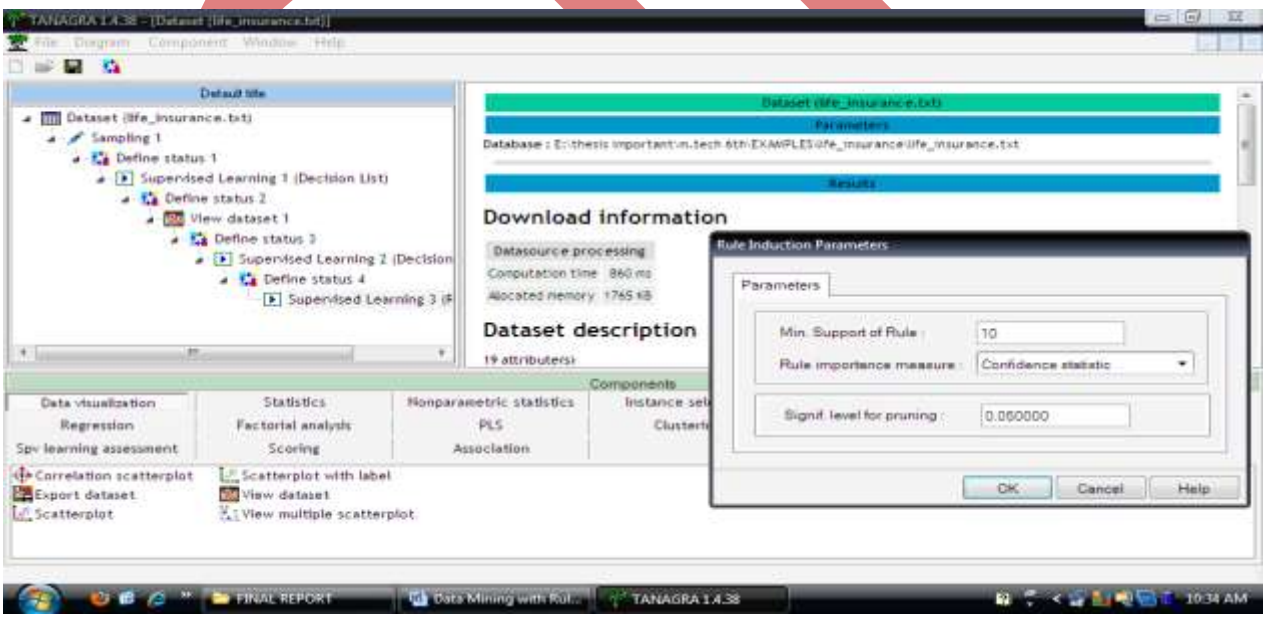


Fig 7

We validate them. We click on the VIEW menu. We obtain **only 1 rule in 250 ms.**

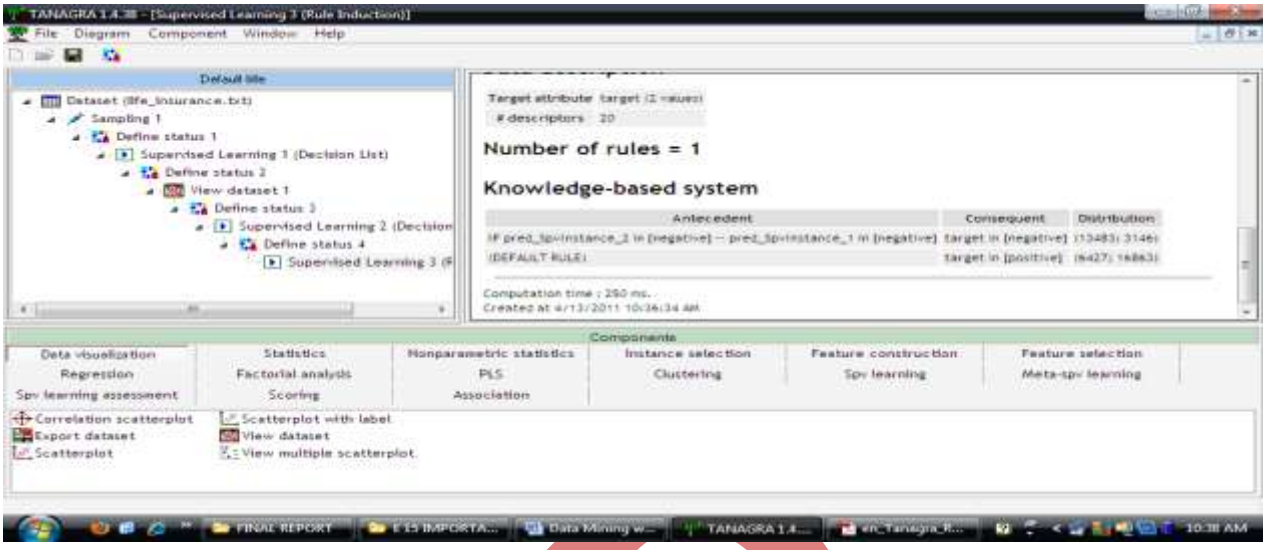


Fig 8

What is the behavior of this classifier on the test sample? We insert a DEFINE STATUS component,

We set “target” as TARGET, the prediction PRED_SPVINSTANCE_2 as INPUT. Then, we add the TEST component. The test error rate is 23.9%. Even if the induction algorithm generates a only one number of rules, they are very relevant.

3.3 Association rule mining:

Apriori Algorithm: Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

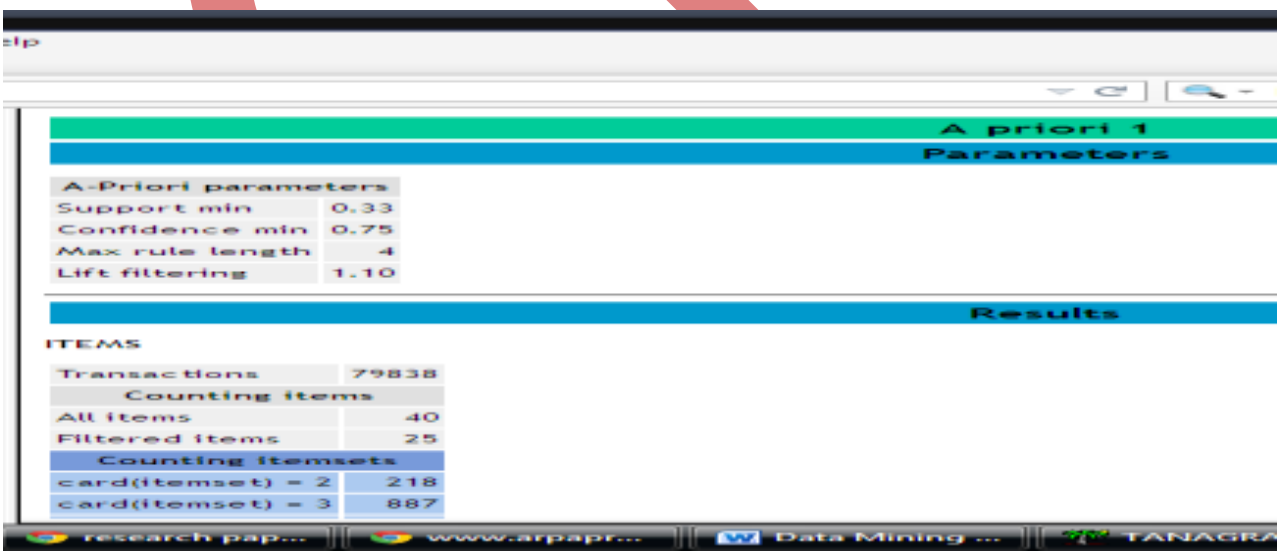
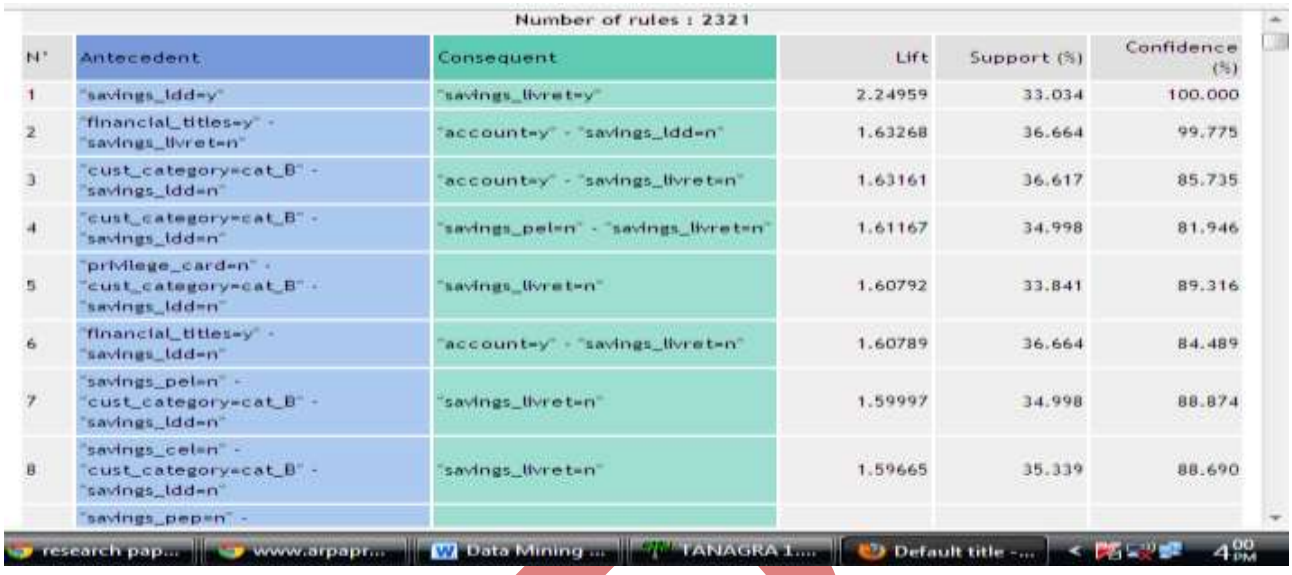


Fig 9: Using Association Rule Mining with min support and min confidence



Number of rules : 2321

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"savings_ldd=y"	"savings_livret=y"	2.24959	33.034	100.000
2	"financial_titles=y" - "savings_livret=n"	"account=y" - "savings_ldd=n"	1.63268	36.664	99.775
3	"cust_category=cat_B" - "savings_ldd=n"	"account=y" - "savings_livret=n"	1.63161	36.617	85.735
4	"cust_category=cat_B" - "savings_ldd=n"	"savings_pel=n" - "savings_livret=n"	1.61167	34.998	81.946
5	"privilege_card=n" - "cust_category=cat_B" - "savings_ldd=n"	"savings_livret=n"	1.60792	33.841	89.316
6	"financial_titles=y" - "savings_ldd=n"	"account=y" - "savings_livret=n"	1.60789	36.664	84.489
7	"savings_pel=n" - "cust_category=cat_B" - "savings_ldd=n"	"savings_livret=n"	1.59997	34.998	88.874
8	"savings_cel=n" - "cust_category=cat_B" - "savings_ldd=n"	"savings_livret=n"	1.59665	35.339	88.690
	"savings_pep=n" -				

Fig 10: shows number of rules

Association rule Mining, Rule Induction Technique and Apriori Algorithm. In Association Rule Mining, we will generate association rules and calculate support and confidence. Assume minimum support and minimum confidence. The rules satisfying both the criteria of minimum support & minimum confidence is true otherwise false. Rule induction technique retrieves all interesting patterns from the database. In rule induction systems the rule itself is of the simple form of "if this and this and this then this". In some cases accuracy is called the confidence and coverage is called the support. Accuracy refers to the probability that if the antecedent is true that the precedent will be true. High accuracy means that this is a rule that is highly dependable. Coverage refers to the number of records in the database that the rule applies to. High coverage means that the rule can be used very often and also that it is less likely to be a spurious artifact of the sampling technique or idiosyncrasies of the database. Assume minimum accuracy and minimum coverage. The rules satisfying both the criteria of minimum accuracy & minimum coverage is true otherwise false.

4 RESULTS

4.1 Error Rate: In Decision list induction, Shannon Entropy supervised algorithms is a best as compare to other algorithms because it has a minimum error rate 24.76%.

4.2 No. of Rules: In rule induction, Misclassification Rate Static supervised algorithm is a best as compare to other algorithms because it has a minimum number of rules is 01.

4.3 Computation Time: In rule induction, confidence static supervised algorithm is a best as compare to other algorithms because it has a minimum computation time is 176ms.

4.4 Using Association Rule Mining:

Error Rate: 33%

No. of Rules: 1815

Computation Time: 11500 ms

	Order Rule Induction Algorithm					
Sampling proportion size 50% Dataset 1	Decision List Induction					
	Supervised Parameters					
	J-Measure			Shannon Entropy		
	Significance	Min Support of rule		Significance	Min Support of rule	
	0.1000	10		0.1000	10	
	Error Rate					
	25.85%			24.76%		
	Value Prediction					
	Value	Recall	1-precision	Value	Recall	1-precision
	Yes	0.7988	0.2300	Yes	0.5645	0.2089
	No	0.8123	0.3246	No	0.8844	0.3211
	No. of Rules					
	24			184		
	Computation Time					
	751 ms			9871 ms		

5 CONCLUSIONS

In this Research paper, we wanted to highlight the approaches for the induction of prediction rules. They are mainly available into academic tools from the machine learning community. We note that they are an alternative quite credible to decision trees and predictive association rules, both in terms of accuracy than in terms of processing time. After analysis Order Rule Induction algorithm is more suitable to find accurate and consuming less access time to mine data with minimum error rate 24.76%.

The theory of Apriori algorithm is that “All nonempty subsets of a frequent item set must also be frequent.” This property prune the candidate which is not in any of the category & thus to reduce number of candidates. I collect the data from Life Insurance Corporation and apply the data mining algorithms to find out the association between the attributes.

REFERENCES

[1] A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D.Malerba. Discovery of spatial association rules in georeferenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6):541–566, 2003.

[2] A. Appice, M. Ceci, and D. Malerba. Mining model trees: A multi-relational approach. In T. Horvath and A. Yamamoto, editors, *Proceedings of ILP 2003*, volume 2835 of *LNAI*, pages 4–21. Springer-V., 2003.

[3] D. Malerba, F. Esposito, M. Ceci, and A. Appice. Top down induction of model trees with regression and splitting nodes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):612–625, 2004.

[4] Globel S., Ranking Mechanisms in Metadata Information Systems for Geospatial Data, *Proceedings of the EOGEO 2002 workshop for developers of Geospatial data services over the Web*, Ispra, 2002.

[5] Johannes Fürnkranz and Gerhard Widmer, Incremental Reduced Error Pruning, in: Proceedings of the 11th International Conference on Machine Learning (ML-94), pages 70--77, Morgan Kaufmann, 1994.

[6] Nocke T. and Schumann H., Meta Data for Visual Data Mining, *Proceedings Computer Graphics and Imaging, CGIM 2002*, Kauai, Hawaii, USA, 2002.

[7] P. Clark and T. Niblett, « The CN2 Induction Algorithm », Machine Learning, 3(4):361:283, 1989.

[8] P. Clark and R. Boswell, « Rule Induction with CN2: Some recent improvements », Machine Learning – EWSL-91, pages 151-163, Springer Verlag, 1991.

[9] In the original CN2 algorithm (Clark and Niblett, 1989), the authors use a more sophisticated beam search during the optimization process. This solution is implemented into the Orange software for instance. The parameter "k" allows specifying the beam width. If we set k=1, we obtain a hill climbing optimization.

[10] http://mydatamining.wordp ess.com/2008/04/14/rule-learneror-rule-induction.

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

Kapil Sharma, Research Scholar, Done B.TECH (CSE) from L.L.R.I.E.T, Moga. Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Data Mining.

+91-8283809270

(Kapilsharma701@gmail.com)

Sheveta Vashisht, Assistant Professor in Department Of CSE, Lovely Professional University, Phagwara, Punjab, India, have done B.Tech, M.Tech from Lovely Professional University, Research area is Networking, Security, Data Mining.

+91-7508280568

(sheveta.16856@lpu.co.in)

Heena Sharma, Research Scholar, Done B.TECH (CSE) from L.L.R.I.E.T, Moga. Now doing M.Tech(CSE) from L.L.R.I.E.T Moga, Punjab, India, Research area is Data Mining.

+91-9464690450

(heenasharma103@gmail.com)

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

Jasreena Kaur Bains, Research Scholar, Done B.TECH (CSE) from Lovely Professional University, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Network security using Neural Network.

+91-8146471880

(Jasreenakaur05@gmail.com)

Richa Dhiman, Research Scholar, Done MSC (IT) from DOABA COLLAGE Jalandhar, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Data Mining, +91-9478285553

(richadhiman58@gmail.com)