International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Comparative Evaluation of Rule Induction Algorithms in Data Mining

Pooja Shrivastava[a*],

[a]Computer Science and Engineering, Jayoti Vidyapeeth Women's University,Jaipur

Accepted 21 May 2013, Available online 1June 2013, **Vol.3, No.2 (June 2013)**

Abstract

Data Mining is a process of extracting useful information from very huge and comprehensive set of data. With the increased possibilities in modern civilization for companies and institutions to gather data inexpensively and efficiently, this subject matter has turn out to be of increasing significance. This interest has stimulated a quickly growing research field with developments both on a hypothetical, as well as on a realistic level with the availability of a range of business-related tools. In this paper we present a comparative result of rule induction algorithms applied on data set of life insurance of persons. This work uses rule induction algorithm in data mining to obtain the precise results with quick processing time. In this paper we are using decision list induction algorithm to make order and unordered list of system to coverage of most data from the data set.

Keywords: *Rule Induction algorithms, data mining, J-Measure, divides and conquers, Shannon Entropy*

1. Introduction

Data mining techniques are the result of a long procedure of research and product expansion. This growth began when industrial data was first stored on computers, sustained with improvements in data Access and more lately, generated technologies that permit users to find the way through their data in real time. Data mining takes this process beyond traditionalist data access and routing to probable and realistic information delivery.

Rule induction is an area of machine learning in which proper rules are extracted from a set of annotations. The rules extracted may symbolize a full methodical model of the data, or merely represent restricted patterns in the data. Using induction rule we can create number of rules for training dataset to accomplish accurate result with a smaller amount of error rate. One can use induction rule algorithms like confidence static and Shannon entropy to get hold of the high rate of precise results from the bulky dataset. This can also improves the conventional algorithms with good result.

Data mining is ready for appliance in the business area because it is supported by three technologies that are now adequately mature:

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

Divide and Conquer Paradigm (A. Appice et al.2003)

In the midst of the rule induction methods, the divide and conquer approach was very popular during mid of 1990's. The Goal is to find out a prediction rule from data
If Premise
Then Conclusion
« Premise » is a set of conditions « Attribute – Relational Operator – Value ».

For instance, Age > 40 and Profession = Workman In the supervised education framework, the feature into the conclusion part is of course the objective attribute. A rule is related to only one value of the target attribute. But one value of the objective attribute may be disturbed by several rules.

Compared to Classification Tree Algorithms

These are based on the divide and conquer concept, their illustration bias is more powerful because it is not constrained by the treelike structure [2]. It sometimes needs a very complicated tree to get an equivalent of a simple rule based system. Some dividing sequences are replicated into the tree. It is known as the replication problem.

Compared to the extrapolative Association Rule Algorithms

They do not suffer because of the redundancy of the induced rules. The plan is even to fabricate the minimal set of rules which allows classifying correctly a new instance. It enables to handle the dilemma of impact about rules, when an occurrence activates two or several rules which

*Correponding author: Pooja Shrivastava

lead to contradictory conclusions. We describe first, divide and conquer algorithms for the rule induction process. Then, we show the performance of the classification rules algorithms implemented by a tool.

Divide and Conquer Algorithms

• Induction of Ordered rules (Decision list induction)

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

The Rule Based System has the subsequent formation [3]

IF Condition x Then Conclusion x
Else If Condition y Then Conclusion y
Else If…
Else If (Default rule) Conclusion Z

Decision List Induction Algorithm

Decision List (Targets , Inputs, Instances)

Ruleset = infinite
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 [3]

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

The classifier is currently outlined as the subsequent structure

If Condition x Then Conclusion x
If Condition y Then Conclusion y
…
(Default rule) Conclusion Z
(Ruleset)

Related work

There are number of handy works that have been presented where most active rule induction algorithms are used. (A. Appice et al.2003) proposed Discovery of spatial association rules in geo-referenced census data.

It is relational mining approach. (F. Esposito et al) proposes Top down induction of model trees with regression and splitting nodes and Ranking Mechanisms in Metadata Information Systems for Geospatial Data. Authors in (J Fürnkranz et al,1994) proposed Rule Induction with CN2 with Some fresh improvements over long-established algorithms. The authors in (Globel S et al,2002) also proposed post and hybrid pruning technique along with rule induction method to obtain high rate of precise results.

They compact the induced set of rules and computational time with high coverage of data from huge data set.

They used decision tree and rule induction method with the help of data mining software.

Solution

Induction of Ordered Rules Dataset

In data set we will take life insurance policy data to detect the customers who are having good policy on the basis of customer categories and to obtain accurate result with less computational time.

Importing the Database

The software used is Tanagra, after the launch of Tanagra, a new diagram is created by clicking on the FILE / NEW menu and import the life insurance .xls file.

Sampling Algorithm

To sub-divide the data-set into a learning sample (50%) and a test sample the SAMPLING Component is being used.

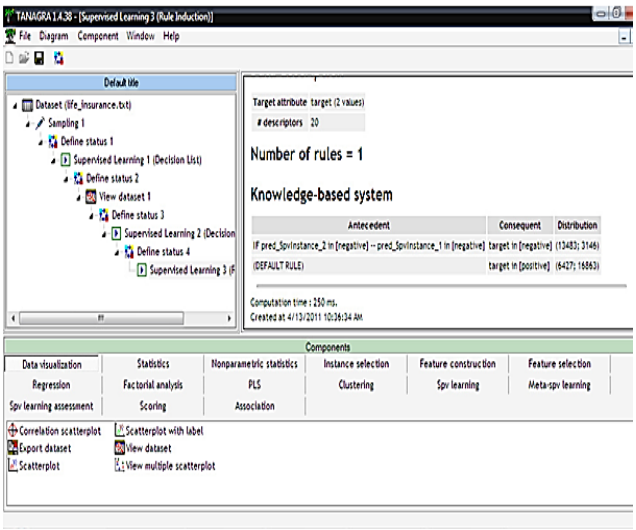

Fig 1: Results of Supervised learning.

Now set target as TARGET attribute and others as INPUT ones using the DEFINE STATUS Component.

Validating these settings and click on the VIEW menu. Result obtained is **20 rules in 703 ms.**

Modify the Parameters of the Learning Algorithm, Validate these settings and click on the VIEW menu. Result obtained is 172 rules in 9203 ms.

Table 1: Comparative results of Rule Induction

	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		

Induction of Unordered Rules Use the RULE INDUCTION component (SPV LEARNING tab) in order to generate a set of unordered rules(K Sharma et al,2013).

Induction of Decision Lists

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

By clicking on the SUPERVISED PARAMETERS menu, the default settings are as it is. Validate them. Click on the VIEW menu. Result obtained is **only 1 rule in 250 ms.**

Insert a DEFINE STATUS component, Settarget 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 an only one number of rules, they are very relevant (K Sharma et al,2013).

Conclusion

The comparative results highlight the approaches for the induction of forecast rules. They are mainly accessible into educational tools from the machine learning area. After analysis Order Rule Induction algorithm is more suitable to find accurate results while consuming less access time to mine data with minimum error rate around 25%, this also improves the conventional algorithms with more precise results, fast processing time and less amount of rules.

Acknowledgement

First of all I would like to thank my mentor Dr. Manoj Shukla for being a great motivator and inspiration for me while writing this paper. He supported me throughout and helped me a lot. After him, I will thank my parents for always being with me. Finally, I would like to thanks my friends for continuously supporting me and helping me throughout.

References

A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D.Malerba (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*.

Appice, D. Malerba and M. Ceci (2003), Mining model trees: A multi-relational approach. In T. Horvath and A. Yamamoto, editors, Proceedings of ILP 2003, volume 2835 of LNAI, *Springer-V*

K Sharma, H. Sharma (Mar 2013), implementation and evaluation of rule induction algorithms in terms of accurate and optimal results in data mining, *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)* ISSN 2249-6831 Vol. 3, Issue 1, 191-200

F. Esposito, D. Malerba, M. Ceci, and A. Appice. Top down induction of model trees with regression and splitting nodes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*

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

Globel S. (2002), 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*.