Rule induction reduces the rule set to maximize the performance

K.Saranya¹, S.Kanimozli²
M.E Computer Science and Engineering
Computer Science and Engineering
Kveri College of Engineering and Technology,
Salem,Tamilnadu

Abstract-The abundance of noisy data in real world gains importance of using rule generation. Pruning in generating classification rules avoids overlapping and over fitting among rules to increase the accuracy and understandability. Rule induction reduces the rule set to maximize the performance of the rule generation algorithm. The proposed pruning strategy amplifies the pruning technique to increase the scalability and clarity among the rules in the final rule set.

1. INTRODUCTION

Recent development organizations, scientific centers produce and accumulate huge amounts of complex and noisy data. To give a reason for their existence and maximize their use, these noisy data need to be stored and analyzed. Because of the information explosion information, data mining has become one of the most important areas where the unsuspected relations between the data are predicted in novel ways that are both understandable and practical to the data owner.

Data mining is a tricky and difficult activity that requires a great deal of expertise in obtaining quality results. So Automation has produces an ever-growing flood of data. Automation of the data mining process requires including many different processing stages in the application design that can deliver the desired results with minimum user intervention. Challenges of Data Mining Scalability Dimensionality/Complexity. Data quality, Data ownership, Privacy considerations, continuously updated data. A common sort of data-mining problem involves discovering unusual events hidden within massive amounts of data.

Data mining is usually used to two main questions are [1] generate prediction based on available data and [2] describe behavior captured in the data. The solution to these queries will be the following approaches are classification, Regression and Time series.

Saranya, Kani,. . . (IJ0SER) June- 2014
3. RULE INDUCTION

Rule learning is an accepted and well researched technique for discovering interesting associations between variables in bulky databases. It is intended to recognize tough rules discovered in databases using dissimilar measures of interestingness. Rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. In data mining, rules are useful for analyzing and predicting customer behavior.

They participate in an important part in shopping basket data study, product clustering, and catalog plan. Each 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 originates in groups with the antecedent. Rule induction is the most established and effective data mining technologies what can be termed 'goal driven' data mining in that a business goal that reduce most comprehensible rules from the given noisy data sets. The objective of applying rule induction data mining is to discover patterns relating the decisions made by the loan underwriters to the details of the application. Rule induction will generate patterns relating the business goal to other data fields (attributes).

The resultant patterns are naturally generated as a tree with splits on data fields and terminal points (leaves) showing the propensity or the magnitude of the business event of interest. Such patterns can reveal the decision making process of the underwriters and their reliability.

Rule Induction is an area of machine learning in which deals with the discovery of high-level, easy-to-interpret classification rules of the form IF-THEN production rules are extracted from a set of observations. Rule induction obtains the most accurate results with fast processing time. Set of rules is useful if rules are not too numerous, clear, and have sufficiently high accuracy.

The rules extracted may represent a technical model, or only represent in local patterns. Rules are used to support decision making in classification (Classification, Machine Learning), regression (Regression, Statistics) and association tasks. In rule based classifiers the rules are composed of two parts mainly rule antecedent and rule consequent.

The rule antecedent, is the if part, specifies a set of conditions referring to predictor attribute values, and the rule consequent, the then part, specifies the class predicted by the rule for an example that satisfies the conditions in the rule antecedent. These rules can be generated using different classification algorithms, the most well known being the decision tree induction algorithms and sequential covering rule induction algorithms.

4. INDUCTIVE LEARNING:

In recent years, there has been a growing amount of research on inductive learning. Induction (or inductive inference) is a method of moving from the particular to the general - from specific examples of general rules. Induction is the process of generalizing a procedural description from presenting or observed examples.

The purpose of inductive learning is to perform a synthesis of novel knowledge, and this is independent of the form given to the input information. In order to form a knowledge base using inductive learning, the initial chore is to gather a set of representative examples of expert decisions. Each example belongs to a known class and is described in terms of a number of attributes. The induction process finds a method of classifying an example, again expressed as a function of the attributes that explains the training examples and that may also be used to classify previously unseen cases. The outcome of an induction algorithm is either a decision tree or a set of rules.

5. PRUNING

One of the classification problems is over fitting that occur regularly and lead to poor accuracy in future predictions. The key to this problem is pruning. Thus the
pruning technique combined with an existing algorithm to improve the performance. Most algorithms use either post pruning or backward pruning or pre pruning or earlier pruning.

1. **Pre pruning**
   Stop generating rules early if the goodness measure is lower than a threshold.

2. **Post pruning**
   Remove the rules from final rule set.

3. **Hybrid pruning**
   Combination of both pre and post pruning.

4. **MINIMUM NEW CLASSIFICATION (MNC)**

   The RULES family of algorithm is also called as covering algorithm. This algorithm differs from the other covering algorithm in the way of inducing rules leads to much more measure of overlapping and over fitting problem. So this RULES family of algorithm makes use of heuristics like H measures, S measures, M measures but it does not provide a good outcome to minimize an overlapping among rules instead it reduces more rules.

   This leads to more uncovered positive examples which degrade the rule set accuracy which makes rule set to be not used in more noisy domain. Coverage refers to the number of seeds that the rule applies to. Calculate MP (the minimum number of positive examples the child rule should cover) and MPS (the minimum support the parent rule should cover). These prevent in generating more rules from the seed examples and also minimize the rules that cover number of negative examples.

   The Minimum New Classification (MNC) a new incremental pre pruning technique attempt with a RULES-7 algorithm to reduce the overlapping problem that is being faced while generating association rules.

   MNC restrict the rules added in ultimate rule set that minimize the correctness of final prediction. MNC restrict the rules added in ultimate rule set that minimize the correctness of final prediction.

   Another major concern of this procedure is marking of rules instead of removing the seed examples after inducing the rule by that it takes the advantage of overcoming fragmentation and disjunct problem. MNC check “the percentage of instance that is being covered by the new rule that has not yet been covered by the rule set created so far”. Using MNC as a criterion, initially RULES-7 procedure makes use of the I_O_R procedure to generate a single rule, at intermediate of I_O_R algorithm the MNC as a threshold to refine each rule while generating. This MNC checks each rule and drop the rule if it does not reach the criterion. If the MNC criterion are reached rule is being marked. Further MNC make use of MNC1 and MNC2 to avoid the problem of computational overhead.

   Only by using minimum coverage as stopping criteria does not prune the rules as better and not able to cover number of positive examples. To make a more accurate rule set the proposed work make use of additional stopping criteria like support and confidence to make the rule set more accurately. In I_O_R additionally minimum support & minimum confidence are as stopping criteria stops adding the fake rules.

   In rule reduction the accuracy is called as confidence and coverage are called as support. Accuracy refers to the likelihood that if the antecedent is true that the precedent will be true. High coverage means that the rule can be used frequently and also that it is less likely to be a false piece of the sampling technique. The rules satisfying both the criteria of minimum accuracy & minimum overage are true otherwise false. Along with MNC the stopping criteria added thus it adds an additional condition to drop the rule and to cover the maximum number of positive examples.

5. **PERMISSIBLE MISCLASSIFICATION (PMC)**

   This incremental post pruning technique is meant specifically to address the problem concern with the noisy data. It works in an approach similar to IREP in that an effort
is made to generalize the rule directly after induction. Difference between IREP and PMC is that the training data are not split into a growing and a pruning set and the evaluation of conditions when a rule is being considered for pruning. Instead, the user specifies a certain acceptable misclassification point with domain knowledge in the form of a threshold referred to as “PMC”. Since the objective in rule induction is to maximize the coverage of the positive examples. So, PMC is defined as “the percentage of the number of negative examples in the seed that the rule should cover to regulate the noise”.

The IREP technique continues to drop conditions greedily until its accuracy on the pruning set decreases. The proposed techniques calculates the ratio rule.Misclassified/rule.Covered for each rule resulting from dropping a single condition from the unique rule and retain the rule for which this ratio is a least. If the misclassified instances are still less than the PMC, it is further generalized. When rule. Misclassified above the PMC level, pruning halts and the final rule for which the number of misclassified instances was less than PMC is chosen as the final rule. A rule is generalized only if rule. Classified is greater than rule. Misclassified otherwise it is pruned away. Rules with a lower coverage level are more level to overfitting the noise than those with a sufficiently high coverage.

8. CONCLUSION:

This paper has presented a hybrid and incremental pruning technique combined with stopping criteria has improved the rule induction process. Thus the computational complexity has been reduced as by making as a pruning user specified. Eventually the speed of an algorithm is being increased.

9. REFERENCE: