ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Framework of Classification Algorithms

M. Amirjahan [1], Dr.N.Sujatha [2]

Research Scholar, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivagangai, TamilNadu, India [1]

Assistant Professor, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivagangai, TamilNadu, India [2]

ABSTRACT: Data mining is a standout amongst the most critical research areas in the field of computer science. Data mining systems are utilized for separating the hidden learning from the extensive databases. There are different search areas in data mining like image mining, data mining, sequential pattern mining, web mining, etc. The reason for data mining is to prepare unstructured data, separate significant numeric files from the data and along these lines make the data contained in the data open to the different data mining algorithms. There are different strategies in data mining, for example, data recovery, report similarity, data extraction, classification, clustering, etc. Looking of comparable archives has a vital part in data mining and document management. Classification is one of the primary assignments in document similarity. It is utilized to classify the archives in view of their class. In this paper, we have investigated the performance of three classification algorithms to be specific Decision Tree Classifier, Rule Based Classifier and Naïve Bayes Classifier. These algorithms are utilized for classifying computer records taking into account of their expansion. The performances of these algorithms are analyzed by employing performance factors like classification accuracy and error rate. From the test results, it is studied that Naïve Bayes performs better than other algorithms.

KEYWORDS: Data mining, classification, Decision Tree Classifier, Rule Based Classifier and Naïve Bayes Classifier.

1. INTRODUCTION

Data mining or Knowledge Discovery from Text (KDT) manages the machine upheld investigation of data. It utilizes techniques from data recovery; data extraction and Natural Language Processing (NLP) Furthermore it connects them with the algorithms and methodologies for Knowledge Discovery of Data (KDD), data mining, machine learning and stats.

Ongoing research in the territory of data mining handles issues of data representation, clustering, classification, or the search and modeling of hidden patterns. [1] Data mining is utilized to portray the use of data mining strategies to robotized revelation of valuable or fascinating learning from unstructured or semi-organized data. Data mining is the method of analyzing so as to integrate the data by analyzing the relations, the patterns, and the operations among textual data semi-structured or unstructured data.

Data mining is otherwise or sometimes called as data content mining alludes to the procedure of getting high-quality data from data. High quality data is ordinarily determined through the divining of some patterns and directions through means like statistical pattern learning. [2] Data mining includes the procedure of organizing the data content inferring designs inside of the organized data lastly assessment and understanding of the result. A portion of the essential uses of data mining incorporate Enterprise Business Word , Data Mining Competitive Intelligence, E-Discovery, \National Security, Intelligence Scientific disclosure particularly Life Sciences, Records Management, Search or Data Access and Social media checking. [3,4]. Some of the technologies that have been created and can be utilized as a part of the data mining procedure are data extraction, concept linkage, summarization, categorization, clustering, subject tracking, data perception and question replying.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

II. CLASSIFICATION TECHNIQUES

2.1 Decision Tree

A decision tree (DT) is a flowchart-like tree complex, body part , where each internal lymph node denotes a examination on an property , each ramification represents an final result of the test, and each leaf node (or terminal node) holds a class label . The topmost node in a tree is the root node. During tree construction, attribute survival of the fittest cadence are used to select the attribute which best partitions the tuples into distinct classes. Three popular attribute selection measures are Information Gain, Gain Proportion , and Gini Indicant When DTs are built, many of the branches may reflect noise or outliers in the training data. [5,6]. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. [7].

Algorithm

ID3(D, Attributes, Target)
1. t = createNode()
2. label(x) = mostCommonClass(D, Target)
3. IF ∀hx, c(a)i ∈ D : c(a) = c THEN return(t) ENDIF
4. IF Attributes = ∅ THEN return(t) ENDIF
5. A∗ = argmaxA∈Attributes(informationGain(D, A))
6. FOREACH a ∈ A∗ DO Da = {(x, c(x)) ∈ D : x|A∗ = a} IF Da = ∅ THEN t 0 = createNode() label(t 0) = mostCommonClass(D, Target) createEdge(t, a, t0) ELSE createEdge(t, a, ID3(Da,Attributes \ {A∗}, Target)) ENDIF ENDDO
7. return(t)

Flow chart

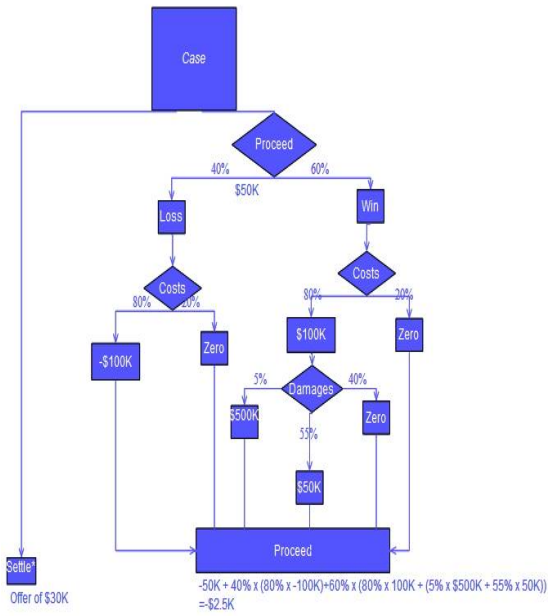

Fig-1 Flow of Decision Tree Classifier

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Advantages

- We can easily understand the DT model.
- Preferences and situation are used to handle hard data using experts.
- To make decision for different situation.

Disadvantages

- Biasing the attributes at different levels.
- Complex calculation of data with uncertain.

2.2 Rule Based Classification

Since data uncertainty is ubiquitous, it is important to develop data mining algorithms for uncertain datasets. We focus on developing a rule-based classification algorithm for data with uncertainty. Rule-based data mining algorithms have a number of desirable properties. Rule sets are relatively easy for people to understand, and rule learning systems outperform decision tree learners on many problems. Rule sets have a natural and familiar first order version, namely Prolog predicates, and techniques for learning propositional rule sets can often be extended to the first-order case. However, when data contains uncertainty - for example, when some numerical data are, instead of precise value, an interval with probability distribution function with that interval - these algorithms cannot process the uncertainty properly. a new rule-based algorithm for classifying and predicting both certain and uncertain data. data . We integrate the uncertain data model into the rule-based mining algorithm. We propose a new measure called probabilistic information gain for generating rules. We also extend the rule pruning measure for handling data uncertainty. We perform experiments on real datasets with both uniform and Gaussian distribution, and the experimental results demonstrate that uRule algorithm perform well even on highly uncertain data. [8].

Algorithm

1. uRule(Dataset D, ClassSet C) begin 1: RuleSet = ∅; //initial set of rules learned is empty
2: for Each Class ci ∈ C do
3: newRuleSet = uLearnOneRule(D, ci);
4: Remove tuples covered by newRuleSet from DataSet
5: RuleSet += newRuleSet;
6: end for; 7: return RuleSet;
End

Flow chart

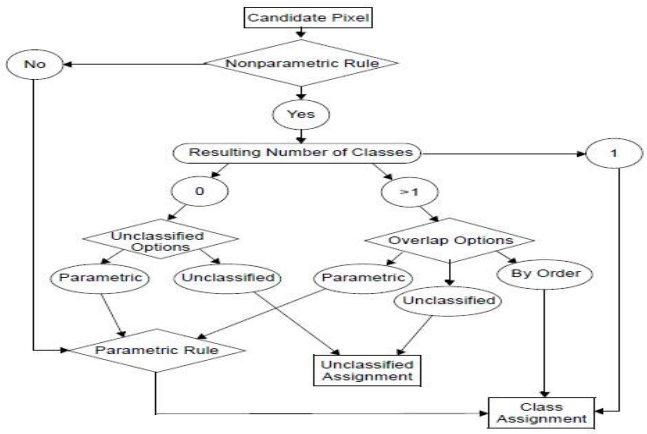

Fig-2 Rule based Classifier

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Advantages

- Simpler, more compact representation
- State is recorded in a memory
  - Create “internal sensors” – Enemy-Recently-Sensed
- Easy to create and understand
  - Can also be represented as rules
    - Decision trees can be learned

Disadvantages

- Decision tree engine requires more coding than FSM
  - Each tree is “unique” sequence of tests, so little common structure
- Need as many examples as possible
- Higher CPU cost - but not much higher
- Learned decision trees may contain errors

2.3 Naïve Bayes

Naïve Bayes is widely used for the classification due to its simplicity, elegance, and robustness. Navie Bayes can be characterized as Navie and Bayes. Navie stands for independence i.e. true to multiply probabilities when the events are independent and Bayes is used for the bayes rule. This technique assumes that attributes of a class are independent in real life. The performance of the Navie Bayes is better when the data set is actual. Kernel density estimators can be used to measure the probability in Navie Bayes that improve the performance of the model. [9,10]. A large number of modifications have been introduced, by the statistical, data mining, machine learning, and pattern recognition communities, in an attempt to make it more flexible, but one has to recognize that such modifications are necessarily complications, which detract from its basic simplicity.

Algorithm

use Algorithm::NBS;
my $ab = Algorithm::NaiveBayes->new; $ab->add_instance (attributes => {foo => 1, bar => 1, baz => 3}, label => 'sports'); $ab->add_instance (attributes => {foo => 2, blurp => 1}, label => ['sports', 'finance']); ... repeat for several more instances, then: $ab->train; # Find results for unseen instances my $result = $ab->predict (attributes => {bar => 3, blurp => 2});

Flow chart

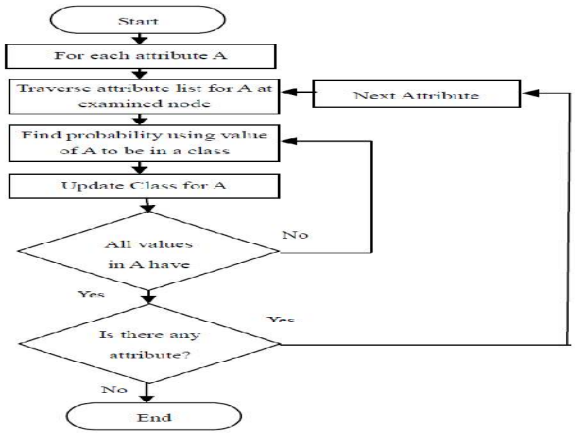

Fig-3 Naïve Bayes Classifier

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Advantages:

1. Fast to railroad train (single scan). CAT scan).

2. Fast to classify – Not sensitive to irrelevant features feature of speech – Handles real and discrete data – Handles streaming data well

Disadvantages:

1. Assumes independence of features

III. EXPERIMENTAL RESULTS

Classification	Decision Tree classifier	Naïve Bayes classifier	Rule based classifier
Accuracy (%)	56.85	92.71	8.71
Time (Sec)	0	0.02	0.02

Table –I: Efficiency Table on Weather Dataset

Classification	Decision Tree classifier	Naïve Bayes classifier	Rule based classifier
Correctly Classified (%)	57.14	64.29	42.86
Incorrectly Classified (%)	42.86	35.71	57.14

Table –II: Classification Table on Weather aset

IV. CONCLUSION

Data mining can be characterized as the extraction of helpful knowledge from extensive data vaults. Data mining is a system which extricates data from both structured and unstructured data and furthermore finding designs which is novel and not known before. In this paper, the Classification algorithms are utilized for characterizing the Weather dataset. The Classification algorithms incorporate three strategies in particular Decision Tree Classifier and Rule Based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Classifier and Naïve Bayes Classifier. By investigating the experimental results the Naïve Bayes Classifier has provide preferable result over other techniques on Weather dataset.

REFERENCES

[1]. Abdullah Wahbeh H, Mohammed Al-Kabi., "Comparative Assessment of the Performance of Three Text Classifiers Applied to Arabic Text", Vol. 21, No. 1, pp. 15- 28, 2012.
[2]. Abdullah Wahbeh H, Qasem Al-Radaideh A, Mohammed Al-Kabi N, and Emad Al-ShawakfaM., "A Comparison Study between Data Mining Tools over some Classification Methods".
[3]. Artur Ferreira., "Survey on Boosting Algorithms for Supervised and Semi-supervised Learning".
[4]. Christophe Giraud-Carrier., "Meta learning - A Tutorial".
[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan-Kaufmann Publishers, San Francisco, 2001.
[6] O. Maimon and L. Rokach, Data Mining and Knowledge Discovery. Springer Science and Business Media, 2005.
[7] X. Niuniu and L. Yuxun, "Review of Decision Trees," IEEE, 2010.
[8] R. Andrews, J. Diederich, and A. Tickle, "A survey and critique of techniques for extracting rules from trained artificial neural networks," Knowledge Based Systems, vol. 8, no. 6, pp. 373–389, 1995.
[9] T. Miquelez, E. Bengoetxea, P. Larranaga, "Evolutionary Computation based on Bayesian Classifier," Int. J. Appl. Math. Comput. Sci. vol. 14(3), pp. 335 – 349, 2004.
[10] M. K. Stern, J. E. Beck, and B. P. Woolf, "Naïve Bayes Classifiers for User Modeling," Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.118.979.