Road Accident Analysis Using Data Mining Techniques

Neha Patil, Prof. Deepesh Jagadale

Abstract– Road accident analysis plays a crucial role within the installation. This paper shows a survey of road accident analysis methods in data processing. within the data processing there's no. of techniques available for clustering and classification, from those techniques k-mean,Apriori algorithm,Naive Bayes classifiers, K Means algorithm. In our existence there aren't any. of accident increases and it's a big problem for us because no. of individuals' death and injuries to improve the road installation is required. During this survey, we then apply Apriori, Naïve-Bayes and K-Means to seek out relationships among the attributes and also the patterns. Naive Bayesian classifier supported Bayes rule is employed to induce the severity.Decision tree classifier is another classifier, which supplies good result for accident severity calculation. Finally K-Nearest Neighbor (KNN) classifier is employed for severity calculation.The accuracy of the algorithms are compared and it's found that KNN performs better than the opposite two algorithms employed. It'll help to enhance analysis accuracy.

I. INTRODUCTION

An algorithm in processing (or machine learning) is also a group of heuristics and calculations that produces a model from data. to create a model, the algorithm first analyzes the data you provide, trying to seek out specific kinds of patterns or trends. The algorithm uses the results of this analysis over many iterations to go looking out the optimal parameters for creating the mining model.

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items within the database and increasing them to larger and bigger item sets as long as those item sets appear sufficiently often within the database.

Naive Bayes classifiers are a set of classification algorithms supported by Bayes' Theorem. it's not one algorithm but a family of algorithms where all of them share a typical principle, i.e. every pair of features being classified is independent of every other.

K Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each datum belongs to just one group. It tries to form the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such the sum of the squared distance between the information points and also the cluster's centroid (arithmetic mean of all the information points that belong thereto cluster) is at the minimum. The less variation we've within clusters, the more homogeneous (similar) the information points are within the identical cluster.

Neural networks represent a brain metaphor for information science. These models are biologically inspired instead of a precise replica of how the brain actually functions. Neural networks are shown to be very promising systems in many forecasting applications and business classification applications because of their ability to "learn" from the information, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize. 2 Procedure for Paper Submission

- Neha Patil *is currently pursuing masters degree program in Msc.IT in Mumbai University ,India,9372304391. E-mail: pneha5055@gmail.com.*
- *Deepesh Jagadale is currently head of IT department in PHCACS, Rasayani in Mumbai University, India, 9028609874. E-Mail: djagdale@mes.ac.in.*

II. Methodology:

1. Apriori Algorithm :

The apriori algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. Apriori uses a " bottom up " approach, where frequent subsets are extended one item at a time a step called candidate generation, and groups of candidates are tested against the information. Apriori is intended to work on databases containing transactions for instance, collections of things bought by customers, or details of a web site frequentation.

Advantages:

- Easy to understand algorithm
- Join and Prune steps are easy to implement on large itemsets in large databases

Disadvantages:

- It requires high computation if the itemsets are very large and the minimum support is kept very low.
- The entire database needs to be scanned.

2. Naive Bayes Classifiers:

This article discusses the speculation behind the Naive Bayes classifiers and their implementation. Naive Bayes classifiers are a set of classification algorithms supported by Bayes' Theorem. it's not one algorithm but a family of algorithms where all of them share a typical principle, i.e. every pair of features being classified is independent of every other.

Advantages of the Naïve Bayes Classifier:

- For problems with a small amount of training data, it can achieve better results than other classifiers because it has a low propensity to overfit. That's Occam's Razor at work!
- Training is quick, and consists of computing the priors and the likelihoods.
- Prediction on a new data point is quick. First calculate the posterior for each class. Then apply the MAP *decision rule*: the label is the class with the maximum posterior.
- The RAM memory footprint is modest, since these operations do not require the whole data set to be held in RAM at once.
- CPU usage is modest: there are no gradients or iterative parameter updates to compute, since prediction and training employ only analytic formulae.
- Scales linearly with number of features and number of data points, and is easy to update with new training data.
- Because of its linear scaling, fast execution, small memory requirement, and light CPU usage, may provide viable solutions for massive problems (many rows and columns) that are too compute-intensive for other methods.
- Easily handles missing feature values — by retraining and predicting without that feature!

Disadvantages of the Naïve Bayes Classifier

- Cannot incorporate feature interactions.
- For regression problems, i.e. continuous real-valued data, there may not be a good way to calculate likelihoods. Binning the data and assigning discrete classes to the bins is sub-optimal since it throws away information. Assuming each feature is normally distributed is workable, but could impact performance if features are not normally distributed. On the other hand, with enough training data in each class, you could estimate the likelihood densities directly, permitting accurate likelihood calculations for new data.
- Performance is sensitive to skewed data — that is, when the training data is not representative of the class distributions in the overall population. In this case, the prior estimates will be incorrect.

3. K-Mean Algorithm:

k -Mean algorithm is used for creating and analysing clusters. In this number algorithm ' n ' number of data points are divided into ' k ' clusters based on some similarity measurement criterion. However results generated using these algorithms are mainly dependent on choosing initial cluster centroids.

K-Means Advantages :

- If variables are huge, then K-Means most of the time computationally faster than hierarchical clustering, if we keep k smalls.
- K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

K-Means Disadvantages:

- Difficult to predict K-Value.
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters.
- It does not work well with clusters (in the original data) of Different size and Different density.

4. Neural Network :

An Artificial Neural Network, often just called a neural network, is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.

Advantages of Artificial Neural Networks (ANN)

- Storing information on the entire network
- Ability to work with incomplete knowledge
- Having fault tolerance
- Having a distributed memory
- Gradual corruption
- Ability to make machine learning
- Parallel processing capability

Disadvantages of Artificial Neural Networks (ANN)

- Hardware dependence
- Unexplained behavior of the network
- Determination of proper network structure
- Difficulty of showing the problem to the network
- The duration of the network is unknown

III. Algorithm :

Road accidents are the main cause of death as well as serious injuries in the world. India is among the emerging countries where the rate at which traffic accidents occur is more than the critical limit. As a human being, everyone wants to avoid traffic accidents and stay safe. In order to stay safe, careful analysis of roadway traffic accident data is important to find out factors that are related to fatal, grievous injury, minor injuries, and non-injury. The relationship between critical rate and other attributes include combining weather conditions, road type, sunlight condi-

tions, speed limit, drunk driver and so on are considered. Here, data mining algorithms are applied on critical accident dataset to address this problem and predict the accident severity. Apriori Algorithm is used for finding an association between attributes. Naive based approach is used for classifying how attributes are conditionally independent. K-means are used to form clusters and analyze them based on attributes. Comparison based on parameters is done to prove the efficiency of the various road accident detection techniques and approaches. The comparison result shows the best road accident detection method. By using these statistics, government/private agencies can take decisions in developing new roads and taking additional safety measures for the general public and awakening a sense of responsibility of road users.

IV. Conclusion :

An algorithm in processing (or machine learning) is also a group of heuristics and calculations that produces a model from data.Apriori is an algorithm for frequent item set mining and association rule learning over relational databases.Naive Bayes classifiers are a set of classification algorithms supported by Bayes' Theorem.K Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each datum belongs to just one group.Neural networks represent a brain metaphor for information science. These models are biologically inspired instead of a precise replica of how the brain actually functions.

V. Reference :

- **Apriori Algorithm:** https://talk.collegeconfidential.com/state-forums/2169222-apriori-is-an-algorithm-for-frequent-item-set-mining-and-association-rule-learning-over-relational-d.html
- https://www.coursehero.com/file/5582021/07apriori/
- **Methodology**
- https://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf
- https://www.vskills.in/certification/tutorial/data-mining-and-warehousing/neural-networks-and-data-mining/
- https://inpressco.com/survey-on-analysis-and-prediction-of-road-traffic-accident-severity-levels-using-data-mining-techniques-in-maharashtra-india/