ISSN 2319 - 6629
Volume 1, No.1, August- September 2012
International Journal of Wireless Communications and Networking Technologies
Available Online at http://warse.org/pdfs/ijwcnt06112012.pdf

Decision Trees for handling Uncertain Data to identify bank Frauds

P.Senthil Vadivu[1], V.Malathi[2]
[1]HOD, Dept. of Computer Applications, Hindusthan College of Arts and Science, Coimbatore-28
sowju_sashi@rediffmail.com
[2]Dept. of Computer Applications, Hindusthan College of Arts and Science, Coimbatore-28
vmalathi89@gmail.com

ABSTRACT

Classification is a classical problem in machine learning and data mining. In traditional decision tree classification, a feature of a tuple is either categorical or numerical. The decision tree algorithms are used for classify the certain and numerical data for many applications. In existing system they implement the extended the model of decision tree classification to accommodate data tuple having numerical attributes with uncertainty described by arbitrary pdfs. So proposed work in this paper is new improved decision tree for both a data representing structure and a method used for data mining and machine learning. The decision trees assist in this work of selecting the attribute that will develop a better performance in finding the required uncertainty information to find the bank fraud.
.

Keywords: Decision Tree, Bank Fraud, Improved New Decision Tree.

1. Introduction

Create a system to classify subscription fraud and bad debts in banking system from large volume of data with minimum false positive alerts. The inputs will be data from a billing database and the outputs will be list of fraudulent characteristics classified as fraudulent or not. The work is focused on discussing how the decision trees are able to assist in the decision making process of identifying frauds by the analysis of information regarding bank transactions. This information is captured with the use of techniques and introduces the new decision tree process. The improved decision tree model of data mining in large operational databases logged from internet bank transactions.

2. REVIEW OF LITERATURE

In this paper we referred survey papers as follows:

Database Mining: A Performance Perspective

In this paper [1], their perspective of database mining as the confluence of machine learning techniques and the performance emphasis of database technology. Their argue that a number of database mining problems can be uniformly viewed as requiring discovery of rules embedded in massive data. They describe a model and some basic operations for the rocess of rule discovery. Their also show how these database mining problems map to this model and how they can be solved by using the basic operations they propose.

Induction of Decision Trees

In this paper [5], focuses on one microcosm of machine learning and on a family of learning systems that have been used to build knowledge-based systems of a simple kind. This paper is concerned with a family of learning systems that have strong common bonds in dimensions. Taking features in reverse order, the application domain of these systems is not limited to any particular area of intellectual activity such as Chemistry or Chess; they can be applied to any such area. While they are thus general-purpose systems, the applications that they address all involve classification.

The product of learning is a piece of procedural knowledge that can assign a hitherto-unseen object to one of a specified number of disjoint classes. Then, the systems described here develop decision trees for classification tasks. These trees are constructed beginning with the root of the tree and proceeding down to its leaves. The family's palindrome name emphasizes that its members carry out the Top-Down Induction of Decision.

Efficient Query Evaluation on Probabilistic Databases: Trees

In this paper [4], they propose such a system. They introduce a new semantics for database queries that supports uncertain matches and ranked results, by combining the probabilistic relational algebra and models for belief. Given a SQL query with uncertain predicates, they start by assigning a probability to each tuple in the input database according to how well it matches the uncertain predicates. Then they derive a probability for each tuple in the answer, and this determines the output ranking.

An important characteristic of their approach is that any query under set-semantics has a meaning, including queries with joins, nested sub-queries, aggregates, group-by, and existential/universal quantifiers2. Queries have now a probabilistic semantics, which is simple and easy to understand by both users and implementers.

Uncertain Data Mining: An Example in Clustering Location Data

In this paper [2], they present the UK-means algorithm, which aims at improving the accuracy of clustering by considering the uncertainty associated with data. Although in this paper they only present clustering algorithms for uncertain data with uniform distribution, the model can be generalized to other distribution (e.g., by using sampling techniques). They also suggest that their concept of using expected distance could be applied to other clustering approaches (such as nearest neighbor clustering and self-organizing maps) and other data miningtechniques (such as data classification)

Efficient Evaluation of Imprecise Location-Dependent Queries

In this paper [3], they study the effect of uncertainty on location-dependent queries, which takes into account the location of the user who issues the query (called "query issuer") in order to decide the query result. An example of such a query is to"find the available cabs within two miles of my current location".

In addition to the uncertainty of the data being queried, the imprecision of the query issuer's location further affects the validity of the query answer. Their goal is to attempt to quantify the query answer validity by efficiently computing the qualification probability of an object for satisfying this type of query i.e., the probability that the object can satisfy this query.

ProTDB: Probabilistic Data in XML

In this paper [6],they would like to model such uncertainty effectively. They would like to insert data into a database even if it is not known with certainty, along with an appropriate indication of the level of uncertainty associated with it. When queried, these certainty levels should be returned with the query results, giving an indication of how likely an element is to satisfy a particular query. Towards this end, probabilistic relational algebras and databases have been proposed by several researchers. Much of the data in the types of applications where uncertainty is an issue, such as web data and scientific data, are not easy to represent in a relational model, even ignoring issues of uncertainty. The flexibility of a semi-structured model is critical, and XML suggests itself as a natural representation choice. Additionally, uncertain information can frequently be represented much more succinctly in XML than in competing relational.

3. EXISTING SYSTEM

In existing system they implement the extended the model of decision tree classification to accommodate data tuple having numerical attributes with uncertainty described by arbitrary pdfs. They have modified classical decision tree building algorithms to build decision trees for classifying such data. They have found empirically that when suitable pdfs are used, exploiting data uncertainty leads to decision trees with remarkably higher accuracies. They therefore advocate that data be collected and stored with the pdf information intact. Therefore, they have devised a series of pruning techniques to improve tree construction efficiency. But this process has some problem when there are tremendous amounts of data tuples are used.

4. PROPOSED SYSTEM

In proposed work we introduce the new improved decision tree process. A decision tree is both a data representing structure and a method used for data mining and machine learning. For this, we can use the concept of a decision tree as a model that maps the observations, taking into consideration a selected attribute as its starting point. The most difficult question here is to find the best attribute. The decision trees assist in this work of selecting the attribute that will develop a better performance in finding the required information. The technique Divide to Conquer is used in decision trees, which consists in breaking the problem into simpler problems, and easier to solve. Now the question is to determine which attribute is going to be the first chosen. We must choose first the ones that have the best information. Decision trees use the concept of entropy to test how much information has an attribute. From the information theory field, we can define entropy as a measure of randomness of a variable. By means of this concept it is possible to measure whether an attribute is explicitly onto a good one. If there are *n* possible messages with equal probability then the probability *p* for each one is *1/n*, and so:

-log (p) = log (n).

Now, for a distribution of **probabilities P = (p1, p2... pn):**

I (P) = - (p1 * log (p1) + p2 * log (p2) + ... + pn * log (pn)).

For example:

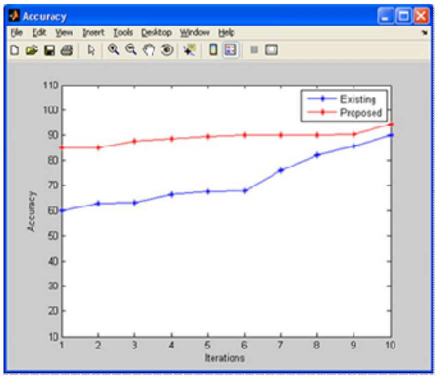
- **If P is (0.5, 0.5) then I(P) is 1.**
- **If P is (0.67, 0.33) then I(P) is 0.92.**
- **If P is (1, 0) then I(P) is 0.**

In these examples, we can see that when the distribution of the probability is higher, then we have better information regarding this variable. This is the basic property at the time of selection of attributes in a decision tree. The entropy is used to estimate the randomness of the variable to predict: the class. The gain of information measures the reduction of entropy caused by partitioning the examples according to the values of the chosen attribute. We define the gain of information as follows:

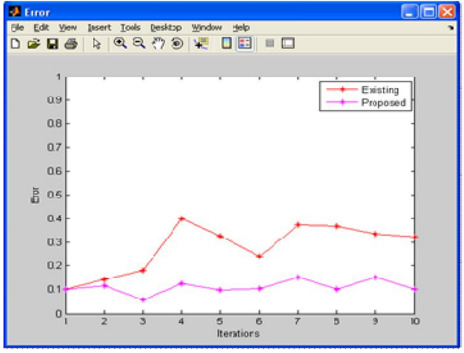
GAIN (attribute) = I (attribute) - I (specific attributes).

The attribute that the GAIN operator shows to provide better information will be used first. The gain of information is used to create small decision trees that can identify the answers with a few questions. The preference is given to the simplest answer according to the Occam's razor principle.

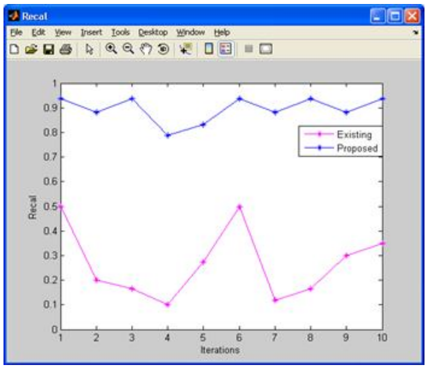
5. EXPERIMENTAL RESULTS



In this above figure we have chosen two parameters called iterations and accuracy which is help to analyze the existing system and proposed systems. The accuracy parameter will be the Y axis and the iterations parameter will be the X axis. The blue line represents the existing system and the red line represents the proposed system. From this graph we see the accuracy of the proposed system is higher than the existing system. Through this we can conclude that the proposed system has the effective accuracy.



In the above figure we are taking the two parameters as error and iterations to analyze the error rate in both existing and proposed algorithm.In X axis the iteration has been taken and in Y axis error parameter has been taken. From this graph we can see that the error rate in existing system is higher than the proposed system. So we can conclude that the proposed system has less error rate which is the best one.



In the above figure we have chosen two parameters called iterations and recal which is help to analyze the existing system and proposed systems on the basis of recal. In X axis the iteration parameter has been taken and in Y axis recal parameter has been taken. . From this graph we see the recal rate of the proposed system is in peak than the existing system. Through this we can conclude that the proposed system has the effective recal.

6. CONCLUSION

Conclusion of this work a decision tree offers the capacity to make classifications, with the help of mathematical concepts, specifically of entropy, studied in information theory, allowing an algorithm to mathematically calculate the randomness of variable regarding the possible choices, thus reducing the difficulty of precisely attaining the goal decision. But, we should be able to create small decision trees so that the goal is reached in a few questions and as quickly as possible. Specifically when the bases are numeric, they can generate huge trees that have a difficult analysis. Scenarios that require quick responses, like bank fraud logs, cannot be used with applications that have a high delay. Based on the implementation of our proposed decision tree, and the test results on a sample real database, we conclude that the decision trees with a criteria for data mining help in decision making, especially in the handling of large data.

REFERENCES

1. Agarawal, T. Imielinski, and A.N. Swami. **Database Mining: A Performance Perspective,** IEEE Trans. Knowledge and Data Eng., Vol. 5, No. 6, pp. 914-925, Dec. 1993.

2. M. Chau, R. Cheng, B. Kao, and J. Ng. **Uncertain Data Mining: An Example in Clustering Location Data,** Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 199-204, Apr. 2006.

3. .J. Chen and R. Cheng. **Efficient Evaluation of Imprecise Location- Dependent Queries,** Proc. Int'l Conf. Data Eng. (ICDE), pp. 586- 595, Apr. 2007.

4. .N.N. Dalvi and D. Suciu. **Efficient Query Evaluation on Probabilistic Databases,** The VLDB Journal., Vol. 16, No. 4, pp. 523-544, 2007.

5.. J.R. Quinlan. **Induction of Decision Trees,** Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986.

6**.** A. Nierman and H.V. Jagadish. **ProTDB: Probabilistic Data in XML,** Proc. Int'l Conf**.** Very Large Data Bases (VLDB), pp. 646-657, Aug. 2002.

7. C.Z. Janikow. **Fuzzy Decision Trees: Issues and Methods,** IEEE Trans. Systems, Man, and Cybernetics, Part B, Vol. 28, No. 1, pp. 1-14, Feb. 1998.

8. T.M. Mitchell. **Machine Learning**. McGraw-Hill, 1997.
.

9. .M. Umanol, H. Okamoto, I. Hatono, H. Tamura, F. Kawachi, S. Umedzu, and J. Kinoshita. **Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems,** Proc. IEEE Conf. Fuzzy Systems, IEEE World Congress Computational Intelligence, Vol. 3, pp. 2113-2118, June 1994.

10. U.M. Fayyad and K.B. Irani., **On the Handling of Continuous- Valued Attributes in Decision Tree Generation,** Machine Learning, Vol. 8, pp. 87-102, 1992.