Construction of Decision Tree for Insurance Policy System through Entropy and GINI Index

Narander Kumar
Department of Computer Science
B. B. Ambedkar University Luck now (U.P.), 226025,INDIA

Vishal Verma
Department of Computer Science,
B. B. Ambedkar University Luck now (U.P.), 226025, INDIA

Vipin Saxena
Department of Computer Science,
B. B. Ambedkar University
Lucknow (U.P.), 226025, INDIA

ABSTRACT
In the modern age of computing, sparse and irregularity in a sample of data is needed for reconstruction of the large and different types of dimensions of data. However, there is a challenge to analyze this type of data. One issue is the robust selection of data. There are many analytical tools are available but crucial part is the decomposition, and make the clusters as well as gradient estimated of data. When one extracts the attributes of sparsely sample data then most likely common attributes may lead to inaccurate results. Due to these, the present paper consists of the solutions through entropy calculation and GINI Index calculation of an insurance company. After calculation of entropy and GINI Index, a decision tree of sample data of an insurance company is presented.

1. Introduction and Related Work

Sparse sampling of data sets is needed to reconstruct a complex and possibly high dimensions of data. If one talks about the scientific simulation, there is a sparse and irregularly distributed sample of data is needed for reconstruction of three or more dimensions of scalar or vector fields in certain spans of interest, without the need to store and process large regular grids. The necessity of sparse sampling arises from the acquisition process. Let us consider an example of study of any company data which is needed to analyze the facts of figures i.e. let us increase the total profit of any company. These aspects are well explained in [1-2]. There are many more techniques which have been developed to analysis and rendering of data on the regular basis. Currently there is an exertion to find efficient corresponding part for irregular and sparse data in low dimensions. However, the shortages of an underlying structure for the samples are techniques and applicable of difficulty to scale to the higher dimensions. The important example is the computation of clusters of data of any company which plays a crucial role to analysis for enhancement of the total profit of the company [3, 4]. The two theorems [5] for presenting the MacDiarmid's bound are explained by authors and one is information gain which is used ID3 (Iterative Dichotomiser 3) algorithm and other is for GINI index, which is used in CART algorithm. They have discussed the technology which is an important for conventional learner decision tree learning system.

Behera [6] has proposed a modified protocol using GINI index over distributed partitioned data which is based on the secure multi party computation for privacy preserving. The author also focused on the algorithm with the problem of making of decision tree and an enhancement of ID3 method for building a decision tree. The reference [7] provides the analysis of rate distraction for dead zone as well as threshold scalar quantization and reconstruction quantization for Gaussian distribution. These analysis are in two parts; first the explanation of the entropy constrained for Gaussian distribution gender the distortion rate function and other one the design criteria for simplify the utilization of effective quantizes. The authors [8] have proposed tree code registration method for group wise alignment of multimodal images using hierarchical intensity space sub division scheme. They computed the accurate estimation of join density function which is based on parzen kernal method. They also proposed a gradient based join entropy minimization scheme to align a group of image. They give comparison between the three code methods with clustering on four different multimodal image dataset. Samui and Samantaray [9] have described the wavelet singular entropy based islanding detection to a micro grid in distributed generation. They processed the three phase voltage signals which are retrieved at target distribution generation location through wavelet transform. After this they derived the coefficient at different levels of decomposition. They generate a matrix of singular value and wavelet singular entropy is also calculated.

Heum et al. [10] have proposed a new complete GINI-Index text feature selection algorithm for text classification. This new algorithm obtains an unbiased feature values and from the feature subsets. This algorithm eliminates many irrelevant and redundant features and also retains many representative features. They also compared the new algorithm with the original versions of algorithm and demonstrated the classification performance. In the reference [11], authors provided classification methods like decision trees, rule mining, Bayesian network, etc. They applied on the educational data for predicting the student behavior, performance in examination etc. To analyze the accuracy of the algorithm, they compared with ID3 algorithm. In [12], authors have proposed the work of predicting the features of dropout student's. To choose the best prediction and analysis, they applied decision tree technique on this study and through which the list of students who are predicted as likely to drop out by data mining is then turn over to teachers and management for direct or indirect intervention. The reference [13] is used for design of decision tree algorithm to classify the students using various criteria. By the use of decision tree one can identify the weak students and a topper through decision tree which helps to improve the result of the class. Decision tree algorithm is used to predict and analysis of the students performance. An algorithm decision tree induction is described in [14]. They take the best partitions the tuples into different classes during attribute selected and construction of the tree. They also discussed about tree pruning and scalability issues for the induction of decision trees from large databases. The computational complexity of the classical decision tree algorithm is described in [15] which is ID3 algorithm. They have proposed an improvement of algorithm

to construct a decision tree with the idea of conditional probability statistical theory.

In the present paper, authors take the clusters of data; compute the entropy and GINI Index for a particular insurance company to evaluate the gradient behavior among the attributes of data associated to insurance company. After calculation of both GINI index and entropy, authors also designed a decision tree which shows the gradient behavior of the edges attributes.

2. Experimental Results

2.1 Computation of Entropy (Information Gain)

ID3 uses the information gain for attribute selection measure and this measure is based on pioneering work by Claude Shannon on information theory, which has the value or "information content" of messages. Let node N represents or hold the tuples of partition D. The attributes with the highest information gain is selected as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. This approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found [16]. The entropy is computed by the following formula:

$$\text{Info(D)} = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (1)$$

Where pi is the probability that a tuple in D belongs to class Ci. Table 1 represents a training sets D, of class-labeled tuples randomly selected from the Insurance Policy System database. The class label attribute, buys policy, has two distinct values (namely, {Yes, No}). Therefore, there are two distinct classes (i.e. m=2). In this paper, we have taken as class C1 corresponds to Yes and class C2 corresponds to No. There are 13 tuples of class Yes and 3 tuples of class No.

Table 1 Class Labeled Training Tuples from the Insurance Policy System Database

S.No.	Age	Policy-income	Incometax_ payee	Policy_ Rating	Policy _Buys
1	Youth	high	Yes	fair	Yes
2	Youth	Low	Yes	Poor	No
3	Youth	Middle	No	Fair	Yes
4	Middle	Middle	No	Fair	Yes
5	Senior	Middle	Yes	Excellent	Yes
6	Middle	Middle	Yes	Fair	Yes
7	Youth	High	Yes	Excellent	Yes
8	Youth	Low	No	Poor	No
9	Middle	High	Yes	Excellent	Yes
10	Senior	Middle	Yes	Fair	Yes
11	Youth	Low	No	Poor	Yes
12	Youth	High	Yes	Excellent	Yes
13	Youth	Low	No	Fair	Yes
14	Youth	Low	No	Fair	Yes
15	Middle	Low	Yes	Excellent	Yes
16	Middle	Low	No	Poor	No
Entropy	**0.2801**	**0.2653**	**0.4132**	**0.6184**	**0.6962**
Gini Index					

To find the Entropy of Policy _Buys, let us put the value from table 1 in equation 1:

$$\text{Info(D)} = -\frac{13}{16}\log_2\frac{13}{16} - \frac{3}{16}\log_2\frac{3}{16}$$

= **0.6962**

To find the Entropy of Age, let us put the value from table 1 in equation 1:

$$\text{Info}_{\text{age}}(\text{D}) = \frac{9}{16}\left(-\frac{7}{9}\log_2\frac{7}{9} - \frac{2}{9}\log_2\frac{2}{9}\right) + \frac{5}{16}\left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) + \frac{2}{16}\left(-\frac{2}{16}\log_2\frac{2}{16} - \frac{0}{16l}\log_2\frac{0}{16}\right)$$

=.4161

Gain (age) = Info(D) - Info_{age} (D) = **0.2801**

To find the Entropy of Income, let us put the value from table 1 in equation 1:

$$\text{Info}_{\text{income}}(\text{D}) = \frac{4}{16}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{5}{16}\left(-\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5}\right) + \frac{7}{16}\left(-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}\right)$$

= .4309

Gain $_{(\text{income})}$= Info (D) - $\text{Info}_{\text{income}}$ (D) = **0.2653**

To find the Entropy of Income tax, let us put the value from table 1 in equation 1:

$$\text{Info}_{\text{income tax}}(\text{D}) = \frac{9}{16}\left(-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9}\right) + \frac{7}{16}\left(-\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7}\right)$$

= 0.2830

Gain $_{(\text{income tax})}$ = Info(D) - Info $_{\text{income}}$(D)

= **0.4132**

To find the Entropy of Policy Rating, let us put the value from table 1 in equation 1:

Info $_{\text{policy rating}}$ (D) =

$$\frac{4}{16}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)$$
$$+\frac{7}{16}\left(-\frac{7}{7}\log_2\frac{7}{7}-\frac{0}{7}\log_2\frac{0}{7}\right)$$
$$+\frac{5}{16}\left(-\frac{5}{5}\log_2\frac{5}{5}-\frac{0}{5}\log_2\frac{0}{5}\right)$$

= 0.0778

Gain $_{\text{(policy rating)}}$ = Info(D) - Info $_{\text{income(D)}}$

= **0.6184**

The following table 2 summarized the above entropy values:

Table 2 Computation of Entropy

E(policy_ buys)	0.6962
E(age)	0.2801
E(policy_income)	0.2653
E(incometax payee)	0.4132
E(policy_ rating)	0.6184

2.2 Computation of Gini Index

The GINI Index measures the impurity of D, a data partition or set of training tuples by using the following formula [16]

$$Gini(D)=1-\sum_{i=1}^{m} p_i^2$$

In order to compute the index of insurance policy system, we take D be the training data from the table 1 where, there are thirteen tuples belonging to the Policy_Buys = Yes and the remaining three tuples belongs to the Policy_Buys = No. A node N is created for the tuples in D. We first use equation (2) for Gini index to compute the drawbacks of the insurance policy system. To find the Gini Index of insurance policy system, we take the value from Table 1.

$$Gini(D)=\left(1-\left(\frac{13}{16}\right)^2-\left(\frac{3}{16}\right)^2\right)$$

= **.34375**

The value of Gini Index from the Table1 = **.34375**

3. Design of Decision Tree

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form in intuive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use of decision tree depends upon the data. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basic of several commercial rule induction systems. Decision tree is learning of decision trees from class-labeled training tuples. A decision tree is a flowchart like tree structure, where each internal node (nonleaf node) denotes a test on attributes, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost nose in a tree is the root node [16].

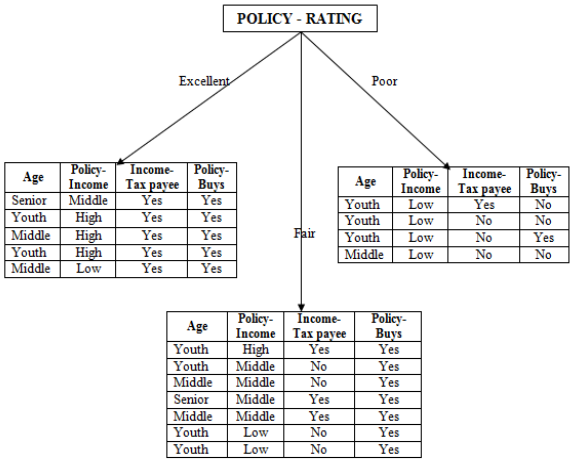

Figure 1. Design of Decision Tree for Insurance Policy System

A decision tree shows in the figure 1, for the concept Policy _Buys, indicates whether a customer at LIC is likely purchase a policy. Each internal (nonleaf) node represents a test on an attributes. Each leaf node represents a class (either Policy _Buys = Yes or Policy _Buys = No). A typical tree shown in the following figure, it represents the concept of Policy _Buys that predicts whether a customer at LIC is likely to purchase a policy. Internal nodes are denoted rectangles, and leaf nodes are also denoted by rectangles. Some decision tree algorithm contains only binary trees, whereas others may be the product of non binary trees.

4. Concluding Remarks

In the present paper, two most popular split functions namely the GINI Index and Information Gain (Entropy) are described with two split parameters Yes/No on the selected split and mathematically characterized values. We analyzed the frequency of Yes/No of GINI index function and the Information (Entropy). Through the entropy result, a decision tree for Insurance Policy System is described which helps the insurance company to select the best rating among the attributes in terms of maximization of the numbers of customers as well as profit gain. Decision tree pruning and scalability issues are the further scope of work.

5. REFERENCES

[1] S. Gerber, P. T. Bremer, V. Pascucci, and R. Whitaker, "Visual exploration of high dimensional scalar

functions." IEEE Transactions on Visualization and Computer Graphics, 16:1271–1280, 2010..

[2] P. Oesterling, C. Heine, H. Janicke, G. Scheuermann, and G. Heyer, "Visualization of high dimensional point clouds using their density distribution's topology." IEEE Transactions on Visualization and Computer Graphics, 99(Pre Prints), 2011.

[3] H. Carr, J. Snoeyink, and U. Axen, "Computing contour trees in all dimensions." Comput. Geom. Theory Appl., 24:75–94, February 2003.

[4] H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci, "Morse-smale complexes for piecewise linear 3-manifolds." In Proceedings of the nineteenth annual symposium on Computational geometry, SCG '03, pages 361–370, New York, NY, USA, 2003. ACM.

[5] Rutkowski, L.; Pietruczuk, L.; Duda, P.; Jaworski, M.; , "Decision Trees for Mining Data Streams Based on the McDiarmid's Bound," Knowledge and Data Engineering, IEEE Transactions on, vol.PP, no.99, pp.1, 0, doi: 10.1109/TKDE.2012.66

[6] Behera, G.; , "Privacy preserving C4.5 using Gini index," Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on , vol., no., pp.1-4, 4-5 March 2011, doi: 10.1109/NCETACS.2011.5751385

[7] Sun, J.; Duan, Y.; Li, J.; Liu, J.; Guo, Z.; , "Rate-Distortion Analysis of Dead-Zone Plus Uniform Threshold Scalar Quantization and Its Application—Part I: Fundamental Theory," Image Processing, IEEE Transactions on , vol.22, no.1, pp.202-214, Jan. 2013 doi: 10.1109/TIP.2012.2215618.

[8] Spiclin, Z.; Likar, B.; Pernus, F.; , "Groupwise Registration of Multimodal Images by an Efficient Joint Entropy Minimization Scheme," Image Processing, IEEE Transactions on , vol.21, no.5, pp. 2546-2558, May2012, doi: 10.1109/TIP.2012.2186145

[9] Samui, A.; Samantaray, S. R.; , "Wavelet Singular Entropy-Based Islanding Detection in Distributed Generation," Power Delivery, IEEE Transactions on , vol.28, no.1, pp.411-418, Jan. 2013 doi: 10.1109/TPWRD.2012.2220987.

[10] Heum Park; Soonho Kwon; Hyuk-Chul Kwon; , "Complete Gini-Index Text (GIT) feature-selection algorithm for text classification," Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on , vol., no., pp.366-371, 23-25 June 2010

[11] S. Anupama Kumar1 and Dr. Vijayalakshmi M.N "Efficiency of Decision Trees in Predicting Student's Academic Performance" Computer Science & Engineering Bibliography, pp. 335–343, 2011.

[12] M. N. Quadri1 and Dr. N.V. Kalyanka- "Drop Out Feature of Student Data for Academic Performance Using Decision Tree", Global Journal of Computer Science and Technology Vol. 10 Issue 2 (Ver 1.0), April 2010.

[13] S.Anupama Kumar, Dr.Vijayalakshmi M.N.,"Prediction of the student recital using classification Technique", IFRSA's International journal of computing (IIJC) , Volume 1, Issue 3, pp305-309, July 2011.

[14] Ravindra Changala, Annapurna Gummadi, classification by Decision Tree Induction Algorithm to Learn Decision Trees from the class-Labeled Training Tuples" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, ,pp 427-434, April 2012.

[15] XianMin Wei, "Research and application of conditional probability decision tree algorithm in data mining" Circuits, Communications and System (PACCS)Second Pacific-Asia IEEE International Conference on, pp 78-80., 1 – 2 Aug 2010

[16] Han J., Kamber M., "Data Mining Concepts and Techniques" Morgan Kaufmann Publishers, San Francisco, CA 94111, ISBN: 978-1-55860-901-3, @ 2006.

AUTHOR'S PROFILE

Narander Kumar received his Post Graduate Degree and Ph. D. in CS & IT, from the Department of Computer Science and Information Technology, Faculty of Engineering and Technology, M. J. P. Rohilkhand University, Bareilly, Uttar Pradesh, INDIA in 2002 and 2009, respectively. His current research interest includes Quality of Service (QoS), Software Engineering, Computer Networks, Resource Management Mechanism, in the networks for Multimedia Applications, Performance Evaluation. Presently he is working as Assistant Professor, in the Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, INDIA.

Vishal Verma is a research scholar in Department of Comupter Science, Babashaheb BhimRao Ambedkar University, and Lucknow, India. Earlier he got his Master of Computer Application (MCA) from the above University and presently he is working on Data Mining Applications through UML.

Vipin Saxena is a Professor and Head, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He got his M.Phil. Degree in Computer Application in 1992 & Ph.D. Degree work on Scientific Computing from University of Roorkee (renamed as Indian Institute of Technology, Roorkee, India) in 1997. He has more than 17 years of teaching experience and 20 years of research experience in the field of Scientific Computing & Software Engineering. He has published more than hundred International and National research papers and authored four books in the Computer Science field. Dr. Saxena is a life time member of Indian Science Congress.