C4.5 Decision Tree Machine Learning Algorithm Based GIS Route Identification

Pankaj Kumar Dalela, Prashant Bansal, Arun Yadav, Sabyasachi Majumdar, Anurag Yadav, Vipin Tyagi
Centre for Development of Telematics Mehrauli, New Delhi, Delhi, India
pdalela@cdot.in, prashant.bansal@cdot.in, arun.yadav@cdot.in, sabyasachi.majumdar@cdot.in, anuragy@cdot.in, vipin@cdot.in

***Abstract*—** Advancement of Geographic Information System (GIS) technologies and Artificial Intelligence (AI) supports development of Decision Support System (DSS) and Spatial Decision Support System (SDSS) to solve complex real life problems. Its areas of implementation encompasses diverse fields from defense to civil departments, from meteorology to disaster management, from traffic analysis to network planning etc. Centre for Development of Telematics (CDOT) has developed a software system for analyzing and identifying optimized GIS routes for laying optical fiber who's planning has been done by Bharat Broadband Network Limited (BBNL). BBNL has done planning for connecting Customer Premise Equipment (CPE) to the Central Office (CO). These routes have been planned by doing manual foot survey. This software system analyses whether the planned GIS routes are optimal or not and suggests optimized fiber routes for planning. In this paper we have presented a solution to the problem that we faced during above mentioned project execution. We have suggested an algorithmic solution based on C4.5 Decision Tree Machine Learning Algorithm. This software system first learns based on user input and then creates rules. Using these rules software system provides fiber planning of optical network over GIS. This algorithm, using machine learning, finds best possible route from a set of routes calculated using different GIS data sources.

Keywords— C4.5; Decision Tree; Machine Learning ; GIS

I. INTRODUCTION

Geographic Information System (GIS) is used to display information over the Earth's surface [1]. Information could be related to a particular area, demographic or geography. Since it is visualized over the earth, it is very easy to interpret and understand. Seeing the capability, many organizations throughout the world want to display information over GIS. Bharat Broadband Network Limited (BBNL), has given mandate to Centre for Development of Telematics (CDOT) of developing a software system that can analyse optical fiber routes over GIS. BBNL has done survey for laying optical fiber and stored fiber length in excel sheets as Survey Report (SR). These fibers are laid between two points, one is village which is termed as Gram Panchayat (GP) where customer premise equipment (CPE) is placed and other is connecting point called Fiber Point of Interconnect (FPOI). These sheets consist Latitude, Longitude of GP and FPOI and the length of optical fiber to be laid between them. So there are two challenges, first one is to verify the length of routes over GIS without going to field and second is to choose best possible route using different GIS data sources.

The verification of GIS route length involves plotting of GP and FPOI on GIS map and finding the best-suited route between them as per available GIS road network as mentioned in [3]. GIS road data plays a significant role in finding optimal routes. GIS road data consists of different layers of road including urban roads, suburban roads, rural roads, village roads, farm roads, private roads. GIS road data received from multiple agencies, differ in coverage and alignment over satellite imagery, hence it becomes very important to consider multiple GIS routes from different sources. In case of urban road network coverage, data from multiple agencies is more or less same, but in other cases, it is different in terms of coverage and quality. It is a time-consuming task to analyse data from different sources and select the best possible route manually. It is found that using only Dijkstra shortest path algorithm to find optimal route length is not enough to cater this problem. Whereas certain others techniques as mentioned in [4], a technique used to generate routes which minimize travel distance, avoid U-Turns and reach higher-priority roads before lower-priority roads can be beneficial. As the process is to be adopted for optical fiber network planning in pan India basis, it becomes very important to achieve a substantial amount of automation supported by decisive capabilities within the machine itself with minimum human involvement. Here Machine Learning Algorithms [2] comes to our rescue, especially C4.5 Decision Tree.

This paper into divided in six sections. In section II we have explained the architecture of software system named Survey Report Analysis (SRA) developed by CDOT. In section III, we have mentioned challenges in finding GIS routes. This section gives details of issues that are faced during algorithm development. In section IV we have discussed about different Decision Tree Machine Learning Algorithms and reason for selecting C4.5 for our solution. Section V talks about development of algorithm based on C4.5 for verifying existing routes and suggesting new routes. Section VI discusses about results that we obtained using C4.5 machine learning algorithm based development and how these positive results can be increased.

II. Architecture of SRA

SRA is designed to perform following features taking inputs received from field survey

- Plotting of GP, FPOI over GIS and finding routes between GP and FPOI over GIS
- Verification of latitude, longitude of GP and FPOI over GIS.
- Verification of route length between GP and FPOI over GIS.

Plotting of GP, FPOI and route between GP and FPOI is done automatically through a client-server model. In this approach the server requests input data from clients and reverts them with the output of SRA. Server goes through numerous route selection algorithms over a number GIS data provided by service providers.

Verification of latitude and longitude of GP and FPOI deals with checking correct position of GP and FPOI. In this feature we test whether the latitude and longitude provided by survey teams are correct or not. This feature enables BBNL to identify fault in location of GP and FPOI position over GIS. In turn BBNL can ask field survey team to recheck latitude and longitude of GP and FPOI.

In verification of GIS routes, we find GIS routes between GP and FPOI and conclude whether these routes are correct by checking them against route length provided in BBNL survey reports.

As shown in figure 1, Survey Reports are first passed through Pre-SRA Quality Analysis (PRE SRA QA), Pre-Data Quality Analysis (PRE DATA QA) phase in which quality check is being done on SR and GIS Data as against standard formats agreed for this task. SRA Client and SRA Server exchanges Input and Output and delivers to Post SRA QA, Output SRA and Data QA module where it is checked for quality.

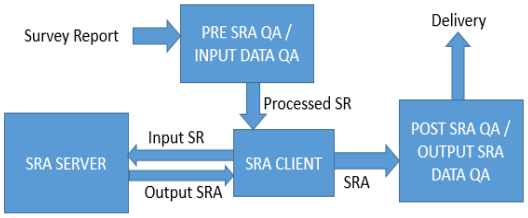

Fig. 1. SRA architecture

III. Challenges in GIS Route Identification

GIS data collection and verification is still a big challenge in rural India. Manual data collection prone to human and technological errors. Due to technological development, accuracy of GPS varies from device to device, apart from technical fault manual fault also adds to the cases. CDOT has been given the task of verifying these GIS routes. CDOT faced a lot of problems while executing this task. Following are the challenges that were faced during verification of GIS routes.

A. *Less Road Coverage*

Fig. 2. Less Road Coverage

Lesser road coverage results in a limited choice of routes available for an algorithm to look for. As one can clearly see there are many roads seen in satellite image but very few are mapped on GIS road layer. This leads to effecting the choice of selecting a correct route. If we consider only one GIS road data source then we may lead to incorrect optical fiber route along the road. Hence there is an urgent need of taking multiple GIS data source into consideration. Our algorithmic approach considers multiple GIS data source for optimal route finding.

B. *Disconnection between roads*

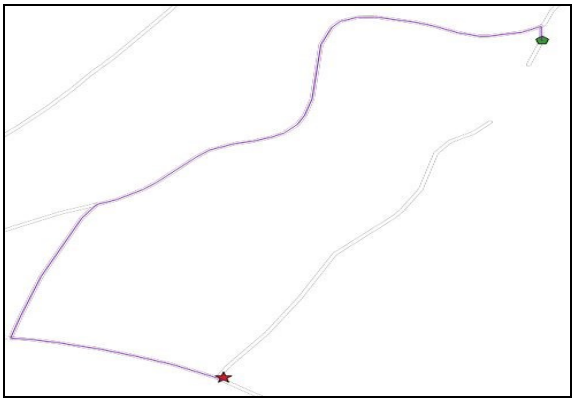

Fig. 3. Disconnection between roads

The disconnection between roads redirects the algorithm to choose next shortest route available between GP & FPOI. As you can see there is a small disconnection between two points (star and dot). This will lead to selection of wrong route. But if you see over satellite, you may find that there is a road between disconnected points. Some other GIS data source may

have these roads as connected, which can be verified by algorithmic approach used in CDOT's system.

C. Misalligned to Satellite Image

Fig. 4. Misalligned to Satellite Image

GIS lengths of misaligned routes do not find a close reference to length of surveyed route and they defer in selection of routes. This could lead to incorrect road length. However not all GIS data have misaligned roads. So there is need to have an approach to select most appropriate GIS data source whose data is aligned to satellite road.

D. Second nearest Problem

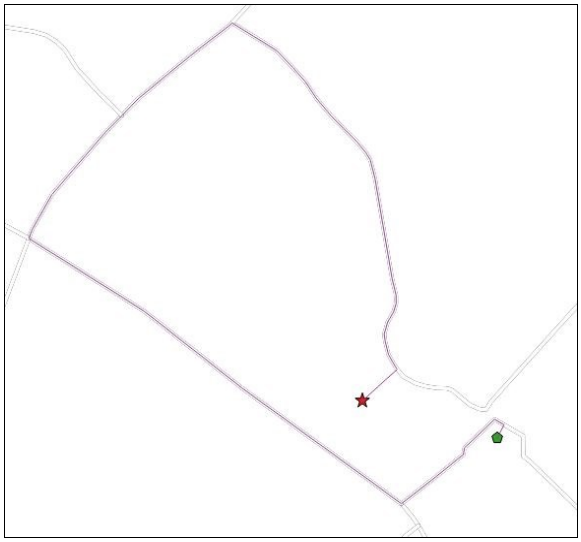

Fig. 5. Second Nearest Problem

Sometimes due to misalignment roads & other irregularities in road data, the GIS route of any service provider may not connect GP and FPOI hence GP & FPOI needs to be connected to nearest node point in the route using line of sight (LOS) distance approach. Second nearest problem forces to connect GP/FPOI to the second nearest node point in route which gives shorter route length as compared to connecting first nearest node point.

For solving these issues and verifying the GIS routes of BBNL's optical fiber, CDOT developed an automated best route finding algorithm based on machine learning.

IV. Decision Tree Machine Learning Algorithm

As mentioned in [6] Decision Tree Machine Learning algorithms are

1. ID3 (Iterative Dichotomiser 3)

   ID3 generates decision tree from a dataset, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set. Higher entropy means improvement in the classification can be done further. This algorithm is used in Machine Learning and Natural Language Processing (NLP).

   ID3 starts with root set S and iterates through every attribute.

   $H(S) = \sum_{x \in X} p(x) \log_2(1/p(x))$

   S – data set for which entropy calculated (for every iteration)

   X – Set of classes in S

   p(x) – The proportion of the number of elements in class x to the number of elements in set S

   When H(S) = 0, all elements in S are of the same class

2. C4.5 (successor of ID3)

   C4.5 is a successor of ID3 algorithm explained above. The decision trees generated by C4.5 can be used for classification, for this attribute C4.5 is often known as a statistical classifier. C4.5 has certain improvements over ID3

- C4.5 handles both continuous and discrete attributes, for continuous attribute, it identifies a threshold and then separate the list.
- Missing attribute are marked as "?" and these attributes are not used in information gain and entropy calculations.
- C4.5 can handle attributes with different weightage
- C4.5 can back trace and delete branches that don't make decisions.

  Since in our algorithmic solution we have used C4.5, it would be better to illustrate C4.5 with example mentioned in [7]. In the figure below, rules have been created providing training data to the algorithm.

Outlook	Temperature	Humidity	Windy	Play (positive) / Don't Play (negative)
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

Fig. 6. C4.5 Training Data

In the table above rules are created based on conditions that are suitable for playing. For instance if it is sunny with temperature and humidity over 85 and winds are still then a rule is created for machine to suggest "Don't Play".

3. CART (Classification And Regression Tree)

   Classification and regression trees (CART) generates either classification or regression trees. CART is a non-parametric decision tree learning technique.

4. CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees [9].

5. MARS: This can extends decision trees so that numerical data can be handled effectively.

6. Conditional Inference Trees. This uses non-parametric tests as splitting criteria. This is corrected for multiple testing to avoid fitting beyond range.

We find C4.5 most suitable for our solution for following reasons

- Due to multiple attribute, it is worth considering C4.5
- Our solution have attributes that have different weightage.
- Some of the attributes may be missing hence algorithm should be able to handle that. C4.5 can be used in this case as well.

V. C4.5 BASED ALGORITHMS FOR GIS ROUTE FINDING

CDOT's algorithmic-based solution has the objective of selecting most appropriate route amongst the routes provided by different GIS data service providers.

Using similar approach as in [8], C4.5 based CDOT's algorithmic-based solution has following steps

1. Training Data for calculation of route paths
2. Rule Generation for comparing route paths
3. Correlation

Training Data

As part of C4.5 training data, we developed an algorithm to create list of data that can be fed to correlation algorithms.

Sample space (G, F) is collected by BBNL doing field survey and provided to CDOT.

G: {set of all GPs}

$\forall\ g_i \in G$: gi consists of pair latitude and longitude

F: {set of all FPOIs}

$\forall\ f_i \in F$: fi consists of pair latitude and longitude.

(g_i, f_i) is an element from set (G,F)

$S_{(gi, fi)}$: surveyed incremental fiber length between (g_i, f_i)

LOS(x,y): function to calculate line of sight length between two GIS coordinates x and y.

Here LOS (g_i , f_i) is line of sight distance between (g_i, f_i)

Assume we have n service providers represented by a set $P= p_1, p_2, p_3 \ldots\ldots\ldots p_n$

The result of algorithm is the decision rule that states which of the service providers have most correct GIS route.

Each service provider provides Rest or SOAP web services to get GIS path between two GIS points.

The path returned from API between two points is represented as $p_i(g_i , f_i)$ and length of fiber route is given by function h(x) i.e. $h(p_i(g_i , f_i))$

Calculated trustworthiness

Also $p_1, p_2, p_3 \ldots\ldots\ldots p_n$ have trustworthiness factor denoted by C=C (p_1), C (p_2), C $(p_3)\ldots$ C (p_n) which is the probability of selection of a particular service provider out of set P.

Such that C (p_1) + C (p_2) + C (p_3) +… C (p_n) =1

C (p_i) = Number of path entries selected from p_i / number of (g, f) processed ……. (formula1)

After each entry of (g_i , f_i) processing, the values of C elements dynamically adjusted to formula 1. This adjusting of values provides a mechanism of correctness and trustworthiness of a particular service provider. Thus, machine learning helps in identifying GIS route which will auto-correct itself based on provided sample space.

Rule Generation

Rules are being generated after training data is fed. These rules are created at the end of this section.

For each entry in (g_i, f_i) in (G, F) route path having greatest of the following factors is calculated and this result is correlated with trustworthiness array.

(1) Common route factor.

(2) Distance correlation with surveyed length and line of sight length.

(3) Road curves.

Common Route factor:
For calculating this factor an upper triangular part pg matrix M of length n x n is formed without diagonal.

To calculate element $m_{k,l}$ of matrix M the route path p_k and p_l are compared as follows:

A path consists of node points. Between p_k and p_l the path with lowest number of node points is picked, suppose here p_l has low number of node points.

$$M = \begin{array}{c|cccccc} & p_1 & p_2 & p_3 & p_4 & \dots & p_n \\ \hline p_1 & & m_{1,2} & m_{1,3} & m_{1,4} & & m_{1,n} \\ p_2 & & & m_{2,3} & m_{2,4} & & m_{2,n} \\ p_3 & & & & m_{3,4} & & m_{3,n} \\ p_4 & & & & & & m_{4,n} \\ \dots & & & & & & \\ p_n & & & & & & m_{n,n} \end{array}$$

Fig. 7. Node Relation Matrix

For each node point x in p_l

If (distance of x to node point in p_k < K (constant))

Then x point considered as matched

Else

x point is unmatched.

Therefore, $m_{k,l}$ is ratio of matched x points to total number of node points.

Similarly all elements of matrix M are calculated.

After this for each element in p_i in P

$\sum P_i = m_{i,1} + m_{i,2} + m_{i,3} + \dots m_{i,n}$ where $n \neq I$

P_i with highest $\sum$ have highest common route factor.

Distance correlation with surveyed length and line of sight length:
As the system has route length of path from n service providers say for (GP, FPOI) g_i, f_i , is $p_1(g_i, f_i)$, $p_2(g_i, f_i)$, $p_3(g_i, f_i)$ …. $p_n(g_i, f_i)$.

Based on above observation, CDOT's algorithm generates following rules.

$\forall$ k in $p_k(g_i, f_i)$ in path array

If ($p_k(g_i, f_i) < LOS(g_i, f_i)$)

then $p_k(g_i, f_i)$ is invalid.

If ($p_k(g_i, f_i) > LOS(g_i, f_i)$ AND

$p_k(g_i, f_i) < S(g_i, f_i)) + K$(variable)

then $p_k(g_i, f_i)$ stands valid.

The $p_k(g_i, f_i)$ with lowest K variable is the most appropriate route on basis of surveyed and line of sight length calculation.

Correlation

Routes p_i calculated from above two methods is being co-related with trustworthiness array C.

For example if route p_1 is found through common factor array and route p_4 is best through distance correlation.

Then C (p_1) and C (p_4) is matched,

If(C (p_1)> C (p_4))

then p_1 route is selected

else

p_4 route is selected.

Finally trustworthiness array is dynamically updated for next (g_i, f_i) processing.

VI. Experimental Results

Above mentioned C4.5 based algorithmic approach is used in finding appropriate routes for 1,30,000 GPs. The results are well satisfactory and saved lots of manual observation time.

Hence we arrive at following conclusions based on our experiment and development

a) Using multiple data sources at input increases data plotting at output

b) Multiple sources at input also gives flexibility of selecting more satellite aligned routes. It helps in creating accurate rules.

c) The average time duration to manually observe a block reduces by 5 times on an average basis.

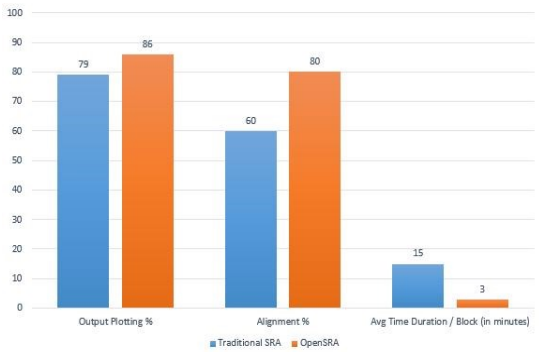

Fig. 8. Effect of using C4.5

VII. CONCLUSION

Above mentioned chart clearly indicates comparison between manual (Traditional SRA which uses manual observation) and C4.5 based SRA (Open SRA which uses machine learning algorithm). Here we see, on average 20 % better output (GIS route plotting) achieved by using C4.5 based solution in comparison to traditional manual approach. More so it has saved 80% time (average time in manual SRA is 15 min and average time in Open SRA is 3 min)

However further research may include image processing, with image processing, we may get even better GIS routes. Image processing can be clubbed with current machine learning based approach and may increase the efficiency of the above algorithm.

REFERENCES

[1] "Concepts and applications of web GIS geo web services-technology and applications", Harish Chandra Karnatak, Indian Institute of Remote Sensing, Indian Space Research Organization, Dehradun, India.

[2] "An introduction to machine learning for students in secondary education", Steven D. Essinger; Gail L. Rosen 2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)

[3] "Geo-intelligence based automatic verification and optimization of manual field survey for OFC network planning" ,Pankaj Kumar Dalela; Saurabh Basu; Anurag Yadav; Sabyasachi Majumdar; Niraj Kant Kushwaha; Arun Yadav; Prashant Bansal; Vipin Tyagi, 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)

[4] Mapping of GPS logs with typical transportation, M. Monica Bhavani; A. Valarmathi,2016 Online International Conference on Green Engineering and Technologies (IC-GET)

[5] Comparing machine learning classification schemes - a GIS approach , A. Lazar; B. A. Shellito,Fourth International Conference on Machine Learning and Applications (ICMLA'05)

[6] Arundhati N., Aamir N. A., Siddharth P., Balwant A. S., "Overview of Use of Decision Tree algorithms in Machine Learning," 2011 IEEE Control and System Graduate Research Colloquium .

[7] http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/c4.5_prob1.html

[8] "The application of decision tree C4.5 algorithm to soil quality grade forecasting model" Li Dongming; Li Yan; Yuan Chao; Li Chaoran; Liu Huan; Zhang Lijuan 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)

[9] Chow, C. K., "A recognition method using neighbor dependence," IRE Transaction on Electronic Computers, vol. 11, 1962,pp.683-690.