Study of Decision Tree Classification Algorithms using Matrimonial System

Pallavi Mude[1], Prof Rahila Sheikh[2]
[1]PG Student, [2]Asst. Professor
[1]Dept. of Computer Science & Eng, Rajiv Gandhi College of Eng. Research & Technology, Chandrapur. India
[2]Department of Computer Technology, Rajiv Gandhi College of Eng. Research & Technology, Chandrapur. India

Abstract— Marriage information has always been an integral part of the knowledge base in any civilized society. Usually some agencies and other de-institutionalized sources become operative in producing and transferring great variety of matrimonial information. In recent time, online matrimony portals accelerate the opportunities of providing newer matrimony services for sharing matrimonial information more comfortably and selectively, though often criticized in terms of adequacy and authenticity of such information. This paper aims to present the matrimonial information system using web data mining. Design and implementation of matrimonial system is based on web data mining which is the application of data mining technique help us to determine pattern from web. The performance of matrimony system uses the C4.5 with bagging and CART decision tree classification algorithm for classifying the perfect match.

Keywords *- Data mining, Matrimonial System, Decision Tree, Web Data Mining, C4.5, CART.*

I. INTRODUCTION

India, a secular democratic republic consisting of 28 states and 7 union territories, has great diversity to an extent perhaps incomparable to any other civilization of the world. It is the second most populous country having more than 1 billion people, 23 official languages with over a thousand dialects, and rich cultures for much of its long history (India, Ministry of I&B, 2009). Racial, ethnic, cultural, linguistic, and religious differences are remarkable where twenty religions flow together including Hinduism, Buddhism, Jainism, Sikhism, Christianity, and Islam. Though all these communities speak different languages, practice different cultures, and observe different social customs, still they pride themselves on being unlike members of the country (Singer & Cohn, 1968). Vast diversity in socioeconomic status pertaining to educational attainment, social power, gender inequality, urbanity, caste, etc. is also evident in India. A landless laborer to billionaire industrialists, tribal illiterates to high-class intellectuals, slum dwellers to NRI and mediocre peoples has received equal attention towards the formation of multifarious groups of the nation. Several other circumstances comprehend the complexities of Indian society over many decades (Singh, 1980). Above all, the India is a nation of unity in diversity. In India, unlike developed countries, information has become inevitable in every sphere of the human society. In fact, any developmental issue to some extent depends on the provision and accessibility of quality information. Now it is being treated as like as marketable commodity. However, phenomenal increase of information sources demands for well-organized systems to make the information accessible pin-pointedly and expeditiously. Such a provision of access has become into reality with the availability of database and network systems exploiting efficient technologies (Simkins, 1983), which have alleviated many ills of information handling activities. Thus, a number of computer-based information systems in different areas have been emerged in India for many years (Literature review, 1990). However the convergence of computer with Internet wrote a dramatic change in accessing effective information, thereby offer us a powerful means of managing information based society (Cronin, 1986). Since last decade, growing interest of social commons toward electronic services, stimulated information-brokers for hosting a number of online matrimony sites in India. Many of them are in operation, and varying in their scope based on vast diversity (viz. racial, ethnic, linguistic, and cultural) in Indian society [1].

Arranged marriages happened in the past either through matrimonial columns in newspapers or through relatives, priests and common friends. People working in India or abroad don't have enough time to look for their match in traditional way. They want to blend old and new sensibilities to find the perfect match of their parent's choice.

Matrimonial system seems to strike a compromise between ancient social traditions and the contemporary attitudes of many people by cutting out the intermediary of arranged marriages. The application is being used to help arrange marriages without relatives or marriage bureaus. Those who have never met or known each other, are culturally different and live diagonally across the globe become life partners.

In this paper we are going to explain in brief the various decision tree classification algorithms like ID3, C4.5, CART, etc. The matrimonial information system working architecture is given below in fig.1.

1) *Data Collection:*

In the data collection and preprocessing web server data base contains two types of data base one is content data base that contain the information like user information and other types of data and second is the server log data base for recording the HTTP transaction (log records). Data collection or data acquisition module collect data from the external web atmosphere to provide resources and material for the latter data mining. From the web environment the data source we get the web pages data, hyperlinks data and history data of user visiting log. Data collection module composed by three independent processes that are data collection, data selection, data search.

2) *Data Preprocessing:*

Data preprocessing mainly renovate and progression the source data acquired in data collection phase and construct the data warehouse of associated themes to generate basic plat- form for data mining process. Data preprocessing is the preparation for data mining and it mainly includes data scrubbing, data integration, data conversion, data reduction, etc. Basically in the data preprocessing step convert the data into the form which is accepted by the data mining algorithm.

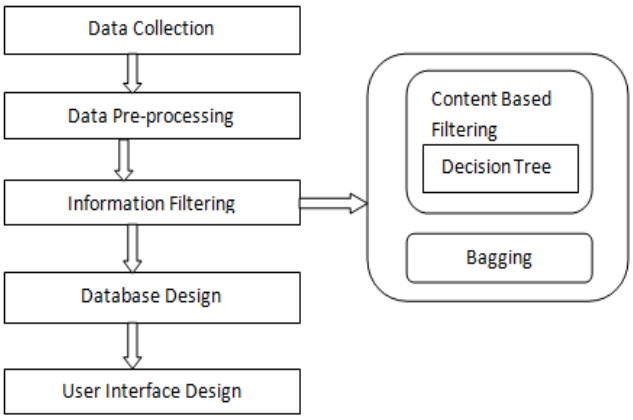

Fig.1. Matrimonial Information System architecture

3) *Information Filtering:*

Information filtering is the main step of the recommendation system. In the existing association rules are applied in the content base filter. In the performance analysis recommendation system, introduced a new architecture based on data mining algorithm for constructing a matrimonial recommender system. The matrimonial recommender system is an intermediary program (or an agent) with a user interface that automatically and intelligently extracts the useful information of bride/groom which suits an individual‘s overall personality. Figure shows the process in the information filter.

a. Content Based Filter: The content-based filtering (CBF) is a consequence and persistence of information filtering research. It constructs the recommendation based on the correlation between difference resources. The output of the content base filter is the suitable match that is search for a person.

b. Decision Tree: Decision rule mining constructs the rule that is applied on user access pattern and generate the result.

c. Bagging: For improving the accuracy of the system we apply bagging on user access pattern.

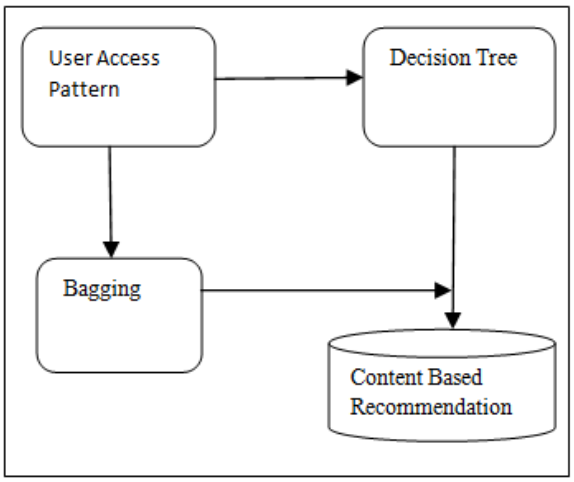

Fig. 2 Information Filtering

4) *Data Base Design and Implementation with Web Based User Interface:*

In matrimonial recommendation system framework, database using the Relational Database Management System (RDBMS) is designed and constructed. This database stores the URLs (i.e., Web pages), keywords for the Web pages, the recommended set of rules from content-based filtering, user login information, and user profiles. MySQL provides a multi- threaded, multi-user, and robust SQL (Structured Query Lan- guage) database management system, which is suitable for the application of recommender systems.

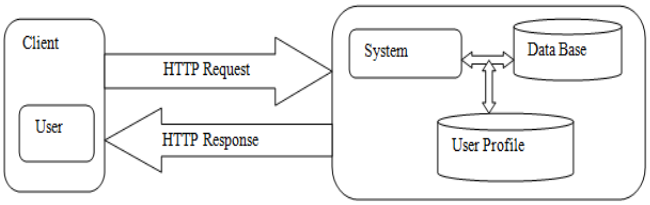

Fig.3 Interaction between user and recommendation system

II. RELATED WORK and LITERATURE SURVEY

Arranged marriages happened in the past either through matrimonial columns in newspapers or through relatives, priests and common friends. People working in India or abroad don't have enough time to look for their match in traditional way. They want to blend old and new sensibilities to find the perfect match of their parent's choice.

Matrimonial system seems to strike a compromise between ancient social traditions and the contemporary attitudes of many people by cutting out the intermediary of arranged marriages. The application is being used to help arrange marriages without relatives or marriage bureaus. Those who have never met or known each other, are culturally different and live diagonally across the globe become life partners.

This literature survey studies about the performance of classification algorithms based on bride/groom data of matrimonial system. The working process for each algorithm is analyzed with the accuracy of classification algorithms. It also studies about various data mining techniques applied in finding the best suitable match for respective bride/groom.

Aman and Suruchi in 2007 have conducted an experiment in WEKA environment by using four algorithms namely ID3, C4.5, Simple CART and alternating decision tree on the students dataset and later the four algorithms were compared in terms of classification accuracy. According to their simulation results, the C4.5 classifier outperforms the ID3, CART and AD Tree in terms of classification accuracy [6].

Nguyen Thai Nghe, Janecek, Haddawy in 2007 presented an analysis on accurate prediction of academic performance of undergraduate and post graduate students of two very different academic institutes: Can Tho University (CTU), a large national university in Viet Nam and the Asian Institute of Technology (AIT), a small international postgraduate institute in Thailand. They have used different data mining tools to find the classification accuracy from Bayesian Networks and Decision tree. They have achieved the best prediction accuracy which is used to find the performance of students. The result of this study is very much useful in finding the best performing students to award with scholarship. The result of this research indicates that decision tree was consistently 3-12% more accurate than Bayesian Network [7].

Sukonthip and Anorgnart in 2011 presented their study using data mining techniques to identify the bad behavior of students in vocational education, classified by algorithms such as Navie Bayes Classifier Bayesian Network, C4.5 and Ripper. Then it measures the performance of the classification algorithms using 10- folds cross validation. It is showed that C4.5 algorithm for the hybrid model yields the highest accuracy of 82.52%. But when it is measured with the F-measure, it is found that the C4.5 algorithm is not appropriate for all data types, but Bayesian Belief Network Algorithm that yields accuracy of 82.4% [8].

Shilpa Dharkar, Anand Rajavat in 2012, proposed a recommendation system which is based on web data mining which is the application of data mining technique help us to determine pattern from web. In terms of accuracy and time performance analysis of recommendation system using two decision tree learning algorithm ID3 and C4.5 and apply it on healthy diet application [9].

T.Miranda Lakshmi, A.Martin , R.Mumtaj Begum, Dr.V.Prasanna Venkatesan in 2013 studied about the performance of classification algorithms based on student data. The working process for each algorithm is analyzed with the accuracy of classification algorithms. It also studies about various data mining techniques applied in finding the student academic performance using ID3, C4.5 and CART decision tree [10].

Anuja Priyam, Abhijeet, Rahul Gupta, Anju Rathee, and Saurabh Srivastava in 2013 proposed decision tree algorithms former applied on the data of students to predict their performance. Performance and results are compared of all algorithms and evaluation is done by already existing datasets. [3].

D. Lavanya, Dr. K.Usha Rani in 2011 presented a paper which was based on performance of decision tree induction classifiers on various medical data sets in terms of accuracy and time complexity are analysed [11].

III. DECISION TREE CLASSIFICATION ALGORITHMS

Decision tree can be constructed relatively fast compared to other methods of classification. Trees can be easily converted into SQL statements that can be used to access databases efficiently. Decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm [2] [3]. The C4.5, ID3, CART decision tree algorithms we are going to apply on our matrimonial data in our next paper to predict their performance. These algorithms are explained below:-

1) *C4.5 Algorithm:*

It is an improvement of ID3 algorithm developed by Quilan Ross in 1993. It is based on Hunt's algorithm and also like ID3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. It accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. It uses gain ratio impurity method to evaluate the splitting attribute.

The algorithm C4.5 has following advantages:

a. Handling attributes with different costs.

b. Handling training data with missing attribute values- C4.5 allows attribute values to be marked as „?" for missing. Missing attribute values are simply not used in gain and entropy calculations.

c. Handling both continuous and discrete attributes- in order to handle continuous attributes,C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.

d. Pruning trees after creation C4.5 goes back through the tree once it has been created and attempts to remove branches that do not help by replacing them with leaf nodes [12].

2) *ID3 Algorithm:*

Iterative Dichotomiser 3 is a simple decision tree learning algorithm introduced in 1986 by Quinlan Ross. It is serially implemented and based on Hunt's algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node [13]. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric – information gain. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked. Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function. ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise and it is serially implemented. Thus an intensive pre-processing of data is carried out before building a decision tree model with ID3.

3) *SPRINT Algorithm:*

It stands for scalable parallelizable induction of decision tree algorithm. It was introduced by Shafer et al in 1996. It is fast, scalable decision tree classifier. It is not based on Hunt's algorithm in constructing the decision tree, rather it partitions the training data set recursively using breadth- first greedy technique until each partition belong to the same leaf node or class. It can be implemented in both serial and parallel pattern for good data placement and load balancing. It uses two data structure: attribute list and histogram which is not memory resident making sprint suitable for large data sets, thus it removes all the data memory restrictions on data. It handles both continuous and categorical attributes [14].

4) *CART Algorithm:*

It stands for classification and regression trees and was introduced by Breiman in 1984.It builds both classifications and regression trees. The classification tree construction by CART is based on binary splitting of the attributes. It is also based on Hunt's algorithm and can be implemented serially. It uses gini index splitting measure in selecting the splitting attribute. CART is unique from other Hunt's based algorithm as it is also use for regression analysis with the help of the regression trees [15]. The regression analysis feature is used in forecasting a dependent variable given a set of predictor variables over a given period of time. It uses many single-variable splitting criteria like gini index, symgini etc and one multi-variable in determining the best split point and data is stored at every node to determine the best splitting point. The linear combination splitting criteria is used during regression analysis. SALFORD SYSTEMS implemented a version of CART called CART using the original code of Breiman. CART has enhanced features and capabilities that address the short-comings of CART giving rise to a modern decision tree classifier with high classification and prediction accuracy.

5) SLIQ Algorithm:

It stands for supervised learning in ques. It was introduced by Mehta et al (1996). It is fast scalable decision tree algorithm that can be implemented in serial and parallel pattern. It is not based on HUNT"S Algorithm for decision tree classification. It partitions a training data set recursively using breadth-first greedy strategy that is integrated with pre-sorting technique during the tree building phase. In building a decision tree model SLIQ handles both numeric and categorical attributes [13].

One of the disadvantages of SLIQ is that it uses a class list data structure that is memory resident thereby imposing memory restrictions on the data. It uses minimum description length principle(MDL) in

pruning the tree after constructing it MDL is an expensive technique in tree pruning that uses the least amount of coding in producing tree that are small in size using bottom-up technique[12].

TABLE 1
Frequency usage of decision tree algorithms
(Algorithm Usage frequency (%))

CLS	9
ID3	68
IDE3+	4.5
C4.5	54.55
C5.0	9
CART	40.9
Random Tree	4.5
Random Forest	9
SLIQ	27.27
Public	13.6
OCI	4.5
Clouds	4.5

Above table 1 shows the comparison between the working of existing algorithms. These algorithms are among the most influential data mining algorithms in the research community [4].

IV. PROPOSED METHODOLOGY

The performance of matrimony system use the C4.5 with bagging and CART decision tree classification algorithm for classify the perfect match. First the content base filters analysis the user access pattern. Content base filter analyzed the user profile whether the user is looking for bride or groom, age , religion, nationality etc are analyzed.

Then according to the user profile matrimony data set is classified by the decision rule mining. It trains the data set and generate rule according to the user access pattern. In matrimony system we use the C4.5 decision rule mining for mining the data and generate rule. These rules are applied on matrimony data set and suggest the suitable match. For performance analysis we calculate the accuracy of the system with C4.5 and then compare the accuracy of C4.5 with CART. For improving the performance of the system we apply bagging with C4.5.

In the performance analysis of matrimonial system decision tree first get the data from content base filter. In the implementation phase we first select the data set then the generated rule. Then these rules are applied into the matrimonial data set. After applying the rule admin selects the profile where we want to apply rule. Once the profile selected the rules are applied and according to the user profile the match is suggested. Then we apply the rules on and analysis the system.

In the performance analysis of matrimonial system decision tree first get the data from content base filter i.e. the training set. In the implementation phase we first select the data set then the generated rule. Then these rules are applied into the matrimonial data set. After applying the rule admin selects the profile where we want to apply rule. Once the profile selected the rules are applied and according to the user profile the match is suggested.

Here we are using two decision tree algorithms i.e. C4.5 and CART. We will apply this two algorithms to matrimony data sets which we get through web and generate the rule. Then these rules are applied to the selected profile. Here we are using C4.5 with bagging to improve the performance. Then the performance analysis of both the decision tree algorithm is analyzed to check the accuracy.

Plan of Work

1) *Input:*

Looking for: Bride / Groom
Age between:
Religion:
Nationality:

2) *Expected Outcome:*

We will give the input like looking for bride or groom, age between, religion and nationality. First it will take the data from web and filter that data as per the requirement of user and then apply the decision tree algorithm which will generate the rule. This rule is applied to the input data set and when there is a perfect match found by the user, it will display the result. We will use both the decision tree algorithms i.e. C4.5 and CART to generate the rule and then the performance analysis is done to check the accuracy. It will be found that the algorithm C4.5 with bagging gives more accurate result than CART.

First the recommendation system suggests the perfect match and then show the comparative analysis of two decision tree classification algorithms in terms of accuracy. For improving the performances of the system bagging is applied. The comparative study of the system shows that after applying bagging it gives more accurate result.

3*) Parameters for performance evaluation:*

Entropy is defined as

$$H(p1, p2, \ldots, ps) = -\text{summation}(pi \log pi)$$

Information Gain is defined as

$$G(D, S) = H(D) - \text{summation } P(Di)H(Di)$$

Gini Index is defined as

$$\text{Gini index} = 1 - \text{summation } P_j^2$$

CONCLUSION

Recommender systems are very popular in the research community, where many approaches have

been proposed for providing recommendations. Recommendation system is used for many application areas. We proposed to use recommendation system in matrimony system.

The matrimony recommendation system uses C4.5 with bagging and CART decision tree classification algorithm for classify the perfect match. First the content base filters analysis the user access pattern. Content base filter analyzed the user profile whether the user is looking for bride or groom, age, religion, nationality etc are analyzed. Then according to the user profile matrimony data set is classified by the decision rule mining. It trains the data set and generate rule according to the user access pattern. In matrimony system we use the C4.5 decision rule mining for mining the data and generate rule. These rules are applied on matrimony data set and suggest the suitable match. For performance analysis we calculate the accuracy of the In our upcoming result paper, we calculate the accuracy of the system by comparing the performance analysis of both the algorithms, i.e.; C4.5 and CART. For improving the performance of the system we apply bagging with C4.5.

REFERENCES

[1] Jiban K Pal , "Review on matrimonial information systems and services – an Indian perspective", International Research Journal of Library, Information and Archival Studies Vol. 1(4) pp. 126-135, November, 2011.

[2] Anju Rathee "survey on decision tree classification algorithms for the evaluation of the student performance" ijct Vol. 4 no. 2

[3] Anuja Priyam, Abhijeet, Rahul Gupta, Anju Rathee, and Saurabh Srivastava, "Comparative Analysis of Decision Tree Classification Algorithms", International Journal of Current Engineering and Technology, ISSN 2277 – 4106, Vol.3, No.2 ,June 2013.

[4] Matthew N. Anyanwu and Sajjan G.shiva "Comparative analysis of serial decision tree classification algorithms", International journal of computer science and security, (IJCSS) Volume 3: Issue (3).

[5] Duan, L., W. N. Street, and E. Xu. Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterprise Information Systems 5 (2), 169-181, 2011.

[6] Aman Kumar Sharma, Suruchi Sahni, " A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering, May 2011 ,Vol. 3 No. 5 ,pp 1890-1895.

[7] Nguyen Thai Nghe; Janecek, P.; Haddawy, P., "A comparative analysis of techniques for predicting academic performance", Frontiers In Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual , pp.T2G-7,T2G-12, Oct. 2007

[8] S.Wongpun, A. Srivihok, "Comparison of attribute selection techniques and algorithms in classifying bad behaviors of vocational education students", Digital Ecosystems and Technologies,2nd IEEE International Conference on , pp.526,531, Feb. 2008

[9] Shilpa Dharkar, Anand Rajavat "Performance Analysis of Healthy Diet Recommendation System using Web Data Mining", International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012.

[10] T.Miranda Lakshmi , A.Martin , R.Mumtaj Begum, Dr.V.Prasanna Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data", I.J.Modern Education and Computer Science, 2013, 5, 18-27 Published Online June 2013 in MECS

[11] D.Lavanya Dr. K.Usha Rani "Performance Evaluation of Decision Tree Classifiers on Medical Datasets", International Journal of Computer Applications (0975 – 8887)Volume 26– No.4, July 2011.

[12] Devi Prasad bhukya and S. Ramachandram , " Decision tree induction- An Approach for data classification using AVL – Tree", International journal of computer and electrical engineering, Vol. 2, no. 4, August 2010.

[13] Tarun Verma, Sweety raj,Mohammad Asif khan, Palak modi, "Literacy Rate Analysis", International journal of science & engineering research volume 3, issue 7, ISSN 2229- 5518. 2012.

[14] Sunita B aher, Mr. LOBO L.M.R.J , "Data mining in educational system using weka tool", International conference on emerging technology trends, 2011.

[15] S.Anupama Kumar and Dr. Vijayalakshmi M.N. , "Efficiency of decision trees in predicting student"s academic performance", D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335-343, 2011.