Towards Judicial Data Mining: Arguing for Adoption in the Judicial System

Owen Kufandirimbwa* and Cry Kuranga

1,2Department of Computer Science, Faculty of Sciences, University of Zimbabwe, Harare, Zimbabwe.

E-mail: kufandirimbwa@gmail.com; Tel: +263712784287
E-mail: ckuranga@cut.ac.zw

INTRODUCTION

The Judicial System collects huge amounts of data which, unfortunately, are not turned into useful information for effective decision making. Decision support system can now use advanced technologies such as On-line analytical processing and data mining to deliver advanced capabilities that ensure efficiency. In this work we developed online analytical processing and data mining model in analysis services using Microsoft Structured Query Language server for judicial system to summaries data, finding the hidden patterns from data for future analysis and prediction, and intuitively present these results to the end users in Microsoft excel (browser). It further exploited decision trees as data mining technique which leads to the design and implementation of the data analysis components to illustrate the feasibility of data mining in the Zimbabwe judicial system.

Keywords: Data mining, OLAP, judicial system, decision tree algorithm, data, information.

Data Mining Requirements and Challenges

Mining abstract and useful information from data can be complicated and sometimes very difficult. The following are requirements and challenges that should be taken into consideration when designing and implementing data mining systems [3].

An effective data mining technique should be robust and powerful. It should have the ability to handle different types of data and performing mining on various types of data structures. Usually a specific data mining system is designed for a specific kind of data because it is impractical to expect a data mining technique to handle all kinds of data.

The searching, mining, or analyzing time of a data mining algorithm should be predictable and acceptable as the size of the database increases [4]. Thus, the performance of the algorithm should degenerate gracefully.

Data mining system should be able to handle noise and exceptional data efficiently. The discovered information must precisely depict the contents of the database and
be beneficial for certain applications. Also, the quality of the discovered information should be interesting and reliable [4]. Data mining facts or conclusions are based on sifting through the data to discover patterns or anomalies [5]. To be effective, the system should allow users to discover information from their own perspectives and the information should be presented to the users in forms that are comfortable and easy to understand [4]. The development of specific graphical user interface is required to express the data mining requests and discovered information.

The Research

Increasingly large amounts of judicial data have been stored and this volume is ever increasing, yet despite this wealth of data, the judicial system has been unable to fully capitalize on its value. This is because information that is implicit in the data is not easy to discern [6].

The main intention is to analyze the feasibility of data mining in the judicial system and ascertain if it’s of benefit. To achieve this aim, we worked to reach three specific objectives:

• To identify a data mining technique that can be used in Judicial Systems.
• To turn low level legal data, usually too voluminous to understand, into higher forms (information or knowledge) that might be more compact (a summary), more abstract (a descriptive model), or more useful (a predictive model)
• To develop a specific user-interface to create data mining model for non technical users.

A Decision Tree (in particular Microsoft Decision Trees algorithm) is a data mining technique employed in this work. According to Microsoft online TechNet library, “the Microsoft Decision Trees algorithm is a classification and regression algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modelling of both discrete and continuous attributes.” Microsoft TechNet also states that, for discrete attributes, the algorithm makes predictions based on the relationships between input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that you designate as predictable, specifically, the algorithm identifies the input columns that are correlated with the predictable column [7].

The decision tree makes predictions based on the tendency toward a particular outcome and for continuous attribute, the algorithm uses linear regression to determine where a decision tree splits [7]. The Microsoft Decision Trees algorithm builds a data mining model by creating a series of splits in the tree and these splits are represented as nodes whereby the algorithm adds a node to the model every time that an input column is found to be significantly correlated with the predictable column [7].

Data required for Decision tree model are [7]:
• A single key column: each model must contain one numeric or text column that uniquely identifies each record. Compound keys are not permitted.
• A predictable column: requires at least one predictable column. You can include multiple predictable attributes in a model, and the predictable attributes can be of different types, either numeric or discrete. However the number of predictable attributes can increase processing time.
Input columns: requires columns, which can be discrete or continuous. Increasing the number of input attributes affects processing time.

DATA EMPIRICAL STUDY: METHODOLOGY

Design of the Proposed Data Mining Architecture for the Judicial System

This is highlighted in figure 1 and its major components
are discussed below:

- **Microsoft Management Console (MMC)**
  MMC is an interface used by administrators to administer, monitor and configure the system [8].

- **Analysis Manager**
  Analysis manager is a tool for the Analysis Server administration in SQL Server Analysis Services which provides user interface for accessing the Analysis servers and the Meta data repositories associated with them [9,10,11]. It is a snap-in application within the Microsoft Management Console (MMC), which communicate with the server through the Decision Support Objects (DSO) component tool [11]. The DSO is a set of programming instructions for applications to work with the Analysis Services [11].

- **ActiveX Data Objects Multidimensional (ADO MD)**
  ADO MD is interface for access to multidimensional schema, query cubes, and also to retrieve the results [12]. It uses an underlying OLE DB provider, which is Microsoft’s strategic low-level Application Program Interface (API) for access to different data sources [13]. OLE DB for Online Analytical Processing (OLE DB for OLAP) is a set of objects and interfaces that extend the ability of OLE DB to provide access to multidimensional data stores [14].

- **PivotTable Service**
  Pivot table service is the primary method for interacting with Analysis Services in order to accomplish such tasks as connecting to a cube or data mining model, querying a cube or data mining model, and retrieving schema information [6]. PivotTable Service also provides methods for online and offline data mining analysis of multidimensional data and relational data [6].

- **Analysis Services**
  Delivers OLAP and data mining functionality for business intelligence applications and supports OLAP by letting you design, create, and manage multidimensional structures that contain data aggregated from other data sources, such as relational databases [15]. For data mining applications, Analysis Services let you design, create, and visualize data mining models that are constructed from other data sources by using a wide variety of industry-standard data mining algorithms including Microsoft Decision tree [15].

- **Olap Cube**
  This is a primary form of data representation within Analysis Services [10]. It is a multidimensional representation of both detailed and summary data, designed according to the client’s analytical requirements. Each cube represents data values of different business entities in which each side presents a different aspect of the data.

- **Cube Browser**
  Cube browser is a layer on top of ADO MD that can be used to write OLAP applications to retrieve data/information from the OLAP Cube.

### Data Collection

The case study (Zimbabwe Civil Court) data consists of approximately 13200 records characterizing different court cases. For each case, the following attributes are provided:

- The applicants details (Name, D.O.B, ID #, Nationality)
- The number of children in the family
- Total monthly income
- Age difference of the applicants in years
- Marriage type- traditional, customary (court) or church based
- Reason for divorce
- Marriage date and divorce
- Judges and the year of the case
- The passed judgment
- Place of stay-province and city
- Applicants educational level
- Applicants occupations
- Applicants smoking or drinking habits
- House-owned couple

None of these has an obvious connection to whether or not the couple is likely to divorce – but data mining can be used to find these connections.

### Data Preparation (Clean the Data)

It is possible to work with the data in its current form, but a few transformations on the data before hand can make more meaningful results. The following transformations were done:

- Male D.O.B – Female D.O.B = Age difference
- Male + Female Income = Total Income
- Date Divorced – Date Married = Married Age
- Couple Nationality [not same = mixed (M), same= same(S)].

### Explanation of Equations

By using the total income attribute as a measure of family income and using the newly created equation instead of the other 8 attributes used in creating them, thus eliminates problems of using the attributes. The equations provide a meaningful comparison of the characteristics of couple even if the couples are of vastly different size. The couple nationality equation aggregates the overall nationality of the couple. After applying these new rules to our dataset, we then create a new dataset (Explored) from our current dataset to work with, excluding those attributes we don’t want (D.O.B, Date Married, Date Divorced, Nationality, Individual Monthly Income).
RESULTS

Relational Data Mining Model using Microsoft Decision Trees

In the Divorce decision tree used, two most important factors that predict the likelihood of divorce is Nationality (mixed marriage - defined by 1st level of the tree) and total number of children (defined by the 2nd level of the tree). For further investigation and navigation of the tree, we can use the content navigator pane i.e. selecting yearly income from prediction column; the decision tree makes this node the root of the current view and creates more space to display all its children (figure 2).

Browsing the Cube using Microsoft Excel

Using a client tool that support PivotTable Service (the Analysis Service Client) we can browse the Analysis Server databases and cubes through a firewall using an internet connection, enhancing data access from anywhere at any time (figure 3). The drilling-down and the drilling-up techniques can also be implemented in Ms Excel.

Graphical representation of data is also possible in Ms Excel, enhancing a deep analytical approach to the mined data thereby leading to much informed decision.

Analysis

Inspection: The results were presented to the judicial expert for validation. The expert ascertains that the decision tree summaries all cases and obtain useful information that can speed decision making, and policy formulation. The results enable the process of analyzing judicial data from different perspectives and summarizing it into useful information, discovering meaningful new correlations, patterns, and trends especially using dependency network.

Accuracy: Before we ran the algorithm we created training and testing datasets. Then we created the model from the training dataset and tested the model on the testing dataset for statistical significance. We split the Explored dataset into two groups of 50%. One group, the Training set will was used as it name implies for training the mining algorithms and the other dataset - Testing was set aside for testing how well our model performs on unseen data. After selecting Total Income as the target attribute and running the Decision Trees, we got the report shown in figure 4.

The report shows that the overall classification probability is 83% - a high accuracy. In addition, it shows low classification errors being: total classification error is 17%; the classification error for "No" is 19%; and the classification error for "Yes" is 13% and determines the
reliability of the Decision Tree model by testing it on unseen data (Data not used in creation of the model). Analysis service will create a report containing the testing results. The reported classification statistics are the same ones used in the original decision trees model. This allows for an easy comparison between the testing results and the original model. Looking at the classification statistics, you will see that the results are only off a few percentiles from the decision tree model report (figure 5a and 5b). For example, the results indicate a Total Classification error around 18%, while the original Decision Tree model had a Total Classification error around 17%. The small deviation between the statistics indicates that the model performance is reliable
Cross Validation: As a benchmark for the designed architecture, a case study that implemented SAS’ SEMMA methodology with Enterprise Miner™ that achieved meaningful, actionable results was used [16].

For the modelling phase, the data were partitioned into 40 percent training, 30 percent validation, and 30 percent testing data sets so that the 20.2/77.8 stratified population distribution was maintained in each partition. The partitioned data set was then sent in parallel into neural network, decision tree, and regression algorithms. Four different configurations of each algorithm were used. All of these algorithms were connected into an Assessment node to evaluate the relative performance of each.

The top decision tree, top regression analysis, and top neural network were connected into an Assessment node. The lift curve is shown in figure 6 [16].

The neural network provided the best predictive power. It is notable that, all three of these models produced significant lift and that the most significant predictors found by the regression and decision tree algorithms were consistent with one another.

Using the above results for validation, we generate Lift charts (figure 7) for the designed architecture to illustrate the effectiveness of mining paradigm. Plotted on the X-
axis is what percentage of all cases that have been acted upon; on the Y-axis is what percentage of attributes which co-relate. The central diagonal indicates a random mining strategy - if you mine 50% of all cases, you will likely mine 50% of the co-relating attributes.

The curve above this axis shows the effect of targeting only those attribute that Analysis Service deemed most likely co-relate in this case study, mining the top 50% of Analysis Service projected cases results in reaching 87% of all the relationships. These results adhere to the lift curves shown in figure 6.

CONCLUSION

The main purpose of the thesis was to establish the feasibility of data mining in the judicial system and how best to implement that technique for non-technical users (judges and lawyers). The Decision Tree model proved to be successful, thus, this would be the model of choice for determining the best possible relationship for the judicial system. The main contributions are summarized as follows:

- Identification of a data mining technique (Decision trees) is appropriate for judicial data set.
- Development of the Data Analysis Components for OLAP solutions is possible.
- Applying the case study of judicial dataset with the user-specific application which implemented the Data Analysis Components was developed in this work.
- Establishment of the remote access to OLAP Cube using Ms Excel was implemented with the Cube Browser.

Recommendation and Future work

Judicial data mining proves to be very feasible and can be of use if implemented in the country's judicial system as a decision support unit for its valuable judicial historic data. DM for domestic security can be considered for future work whereby DM can be used as the technology for identifying links or for developing descriptive or predictive model. This usually leads to three discrete applications for automated analysis which are: subject-oriented link analysis, pattern-analysis and pattern-matching. The policy question then becomes not one of what technology is employed but one of specific application - that is, what data is it permissible to access, using what methodology and for what purpose.

REFERENCES


