Information Decay in Building Predictive Models Using Temporal Data

Lasheng Yu
Department of Computer Science ,Central South University, Changsha, China, 410083
Email: ley462@163.com

Mukwende Placide
Department of Computer Science ,Central South University, Changsha, China, 410083
Email: mukwende@gmail.com

Abstract—Most predictive data mining methods are based on the assumption that the historical data involved in building and verifying a model are the best estimator of what will happen in the future. However, the relevance of past to the future depends on the application domain in a specific timeframe. In this paper, we present a method of predicting the future from temporal data using information decay technique that estimate the relevance of the past to the future with the help of the knowledge of the information lifetime of the temporal data. We show how to use a decision tree data mining technique to build an information-decay-based predictive model by introducing an information decay method in impurity measure functions. The relevance of the concepts is proven by comparing two decision tree-based classifiers: one built using information decay over a specific timeframe and the other built setting aside the time factor.

***Index Terms*—Information decay, predictive model, data mining, classification, temporal data**

I. Introduction

Things change faster than ever. With the advance of technology, every so often, a new technology comes along and changes the way we live our daily lives; things are never the same again. It is not a surprise to see a person changing his behavior over time. For instance, in a supermarket, the characteristics of customers needing an affinity card or purchasing a particular product may vary with respect to time. Time is an important aspect of all real world phenomena. Data record facts or event of our interest; and in most application areas that have been studied for data mining, the time that something happened is also known and recorded. In real life, the temporal data can be found everywhere, such as corporate transactions, customer histories and demographic information, process control data, credit bureau information, weather data, etc. Therefore, any system approaches or techniques that are concerned with knowledge discovery need to take into account the temporal aspect of data [1]. However, the traditional data mining techniques tend to treat temporal data as best as unordered collection of events or facts, and they ignore time values, which the data are stamped with.

Temporal data mining [2] is an extension of traditional data mining as it has the capability of mining activity rather than just states and, thus, inferring relationships of contextual and temporal proximity, some of which may also indicate a cause-effect association. In particular, the accommodation of time into mining techniques provides a window into the temporal arrangement of events and, hence, the ability to suggest cause effect that are overlooked when the temporal component is ignored or treated as a simple numeric attribute

Some researches [1], [2], [3] assessed the advantages of using time value in mining temporal data; frameworks and representational languages [1], [3] for mining temporal data have been developed; change detection methods (in temporal and time series data) have been discovered [4]; knowledge decay or information decay theories remain among the discussion topic in data mining and data warehouse researches [5], [6]. The question stay on how to apply these methods and concepts in existing data mining technique to improving the performance of resulting models for better prediction of the future facts or events.

The work presented in this paper pays attention on how to weigh (using an information decay function), over a specific timeframe, the information given by a case or record during the model building. We start with a discussion of the problems of traditional data mining on temporal data; we then look at the information decay in chronological data and its relevance in predicting the future. A method of building a predictive classifier based on decision tree mining technique is presented. To evaluate our theoretical concepts, we show how we built and tested two models using 940 temporal records collected from SH schema of Oracle database. Using the same build data and parameters and the same test data and parameters, we prove how the first model that is built using time-based information decay is a better predictor of the future than the second model that is built ignoring the time factor. A comparison between the two mining techniques is made in results section where a Receiver Operating Characteristic (ROC) Curve, prediction accuracy, and F1-measure test metrics have been used. Lastly, we conclude and highlight future work.

II. Challenges of Traditional Predictive Data Mining on Temporal Data

The two high-level primary goals of data mining, in practice, are prediction and description. Prediction involves using some variables or fields in the data to predict unknown or future values or other variables of interest. Description focuses on finding human-interpretable patterns describing the data. The predictive nature of data mining is that the models produced from historical data have the ability to predict outcomes such as which customers are likely to churn, who will be

doi:10.4304/jsw.7.2.479-484

interested in a particular product, which medications are likely to affect the outcomes of cancer treatment positively. Two types of predictive models exists: classification models—predict into what category or class a case or record falls, for instance, which offer a customer will respond to; regression models—predict what number value a variable will have, for example, what is the predicted value of a home or a person's income. During predictive model building, historical data is split into two datasets: one for building the model, and the other for testing it. Test dataset is typically not used to build a model in order to give a true assessment of a model's predictive accuracy [7]. There is no restriction on proportion of data used for training (building) and testing; it is up to the data miner to find out [7], [8]. In addition, no restriction is placed on the ordering of records, based on their chronological order, during the historical data split.

During model building and testing, we tend to believe [1], [4], [6] that the mass of data, incrementally accumulated into databases over time, resembles to data that represents future facts or events. However, things change. Consider, for example, a particular supermarket wants to predict which of its customers are likely to increase spending if given an affinity card. The supermarket could build and test a model using demographic data about customers who have used an affinity card in the past, assuming that a positive class (target value of 1) is assigned to customers who increased spending with an affinity card, and a negative class (target value of 0) is assigned to customers who did not increase spending. Suppose that a decision tree algorithm is used to build the model and generate a rule that states that married customers who have a college degree (Associates, Bachelor, Masters, Ph.D., or Professional) are likely to increase spending with an affinity card. If the supermarket used historical data collected, let's say for the past ten years, that rule might be, with a certain confidence and support, true. If, however, the highest concentration of customers who are married and have a college degree (Associates, Bachelor, Masters, Ph.D., or Professional) can be found up to the first five years ago, then the discovery of the rule is not significant for the present and the future of the supermarket. The existing approaches, therefore, take into account all available data (or a subset of them) and provide a static view of an application domain. Application domains, however, do change over time; customers change their habits and things that may have been true in the past are not applicable any more. Temporal data mining can provide a model of an evolving business domain. It can uncover patterns that are applicable during certain time periods. If data mining is to be used a vehicle for better decision making, the existing approaches will in most cases lead into not very significant or interesting results [1]. Therefore, we discuss in subsequent sections our new approach of building predictive models using information decay method. We start explaining the concepts of information decay, followed by its application in a decision tree data mining approach.

III. Information Decay Concepts

A. Temporal Data Definitions

P. Tan, M. Steinbach, and V. Kumar [8] define temporal data as an extension of record data, where each record has a time associated with it. The extension to this definition is where a time may also be associated with each attribute of a record. This definition is applicable, for instance, to retail transaction dataset that also stores the time at which the transaction took place.

F.Roddick, and M. Spiliopoulou [2] categorized temporal data based on temporality aspect of data:

- **Static**: No temporal context is included and none can be inferred.
- **Sequences**: Ordered lists of events. This category includes ordered, but not timestamped collections of events or facts.
- **Timestamped**: A timed sequence of static data taken at more or less regular intervals.
- **Fully temporal**: Each tuple in a time-varying relation in the database has one or more dimensions of time.

In this paper, we define temporal data as a dataset of cases (records) where each case has an associated timestamp that represent the time relationship between the predictors and target attributes. Consider a dataset D of size n (number of records: $R_1, R_2, \ldots, R_i, \ldots, R_n$) and of dimension m (number of attributes: $A_1, A_2, \ldots, A_j, \ldots, A_m$). If we let v_{ij} be the value of j^{th} attribute of i^{th} record, then D is represented in relational manner as shown in figure 1. If D_T is temporal with the j^{th} attribute $T{=}A_j$ (having values $t_1{=}v_{1j}$, $t_2{=}v_{2j}$, $t_i{=}\ v_{ij}$, *and* $t_n = v_{nj}$) as the timestamp, then it is represented relationally as shown in table 1. In other words, if A_k is the target attribute, the temporal data D_T that has timestamp attribute T is a dataset of records: $R_1, R_2, \ldots, R_i, \ldots, R_n$ and attributes: $A_1, A_2, \ldots, T, \ldots, A_m$, where the attribute T define a time relationship between predictors :$A_1, A_2, \ldots, A_m$ (all attributes excluding T and A_k) and the target attribute A_k.

B. Information Decay

While processing temporal data, it is common to down-weight records that correspond to older data [6]. Intuitively, this reflects the belief that most recent data are most relevant for predicting the future, while older data are of less significance, and can be counted at a lower weight or ignored entirely. G. Cormode, V. Shkapenyuk, D. Srivastava, and B. Xu [6] define a decay function as a function that takes the timestamp t_i of record R_i, and returns a weight $f(i,t_i)$ for this record. If t_p is the present time, the function $f(i,t)$ is therefore a decay function if it satisfies the following properties:

1. $f(i,t) = 1$ when $t = t_p$ and $0 \le f(i,t) \le 1$ for all $t \le t_p$.
2. $f(i,t)$ is monotone nonincreasing as time increases: $t \ge t_p$ implies that $f(i,t) \le f(i,t_p)$.

Various decay functions that are widely in-use use the age of a record, instead of timestamp, as the decay function parameter. If t is a timestamp of a record, then the record's age a is computed as $a = t_p - t$, and the decay function is written as $f(a)$. The most popular decay functions that have been used for many applications are:

Table 1. Relational representation of temporal data

A_1	A_2	…	T	…	A_m
v_{11}	v_{12}	…	t_1	…	v_{1m}
v_{21}	v_{22}	…	t_2	…	v_{2m}
…	…	…	…	…	…
v_{n1}	v_{n2}	…	t_n	…	v_{nm}

- **No Decay**: $f(t) = 1$, all records are equally treated regardless of their timestamp information.
- **Sliding Window**: $f_W(a) = 1$ for $a < W$ and $f_W(a) = 0$ for $a \geq W$, where W is the window size parameter. In this case W is regarded as a timeframe in which records are somehow relevant to the future for better prediction of the future.
- **Exponential Decay**: $f(a)=(1 - \lambda)^a$ or $f(a)= exp(-\lambda a)$ for $\lambda > 0$, where λ is the decay constant or simply decay rate. If data is estimated to have a mean lifetime τ, then $\lambda = 1/\tau$. Estimating τ depends on the application domain and the knowledge of the data miner on temporal data. It is therefore subjective.
- **Polynomial Decay**: $f(a) = (a + 1)^{-\alpha}$. This function is required because is slower than the exponential decay, which is too fast.

In this paper, we use a modified version of Sliding Window by letting the function f_W exponentially decay for $a<W$ where $W=\tau$; the resulting function is $f_\tau(a)= exp(-a/\tau)$ for $0\leq a <\tau$, $f_\tau(a)=0$ for $a \geq \tau$. The mean lifetime τ can be calculated from the domain knowledge of the dataset by taking the difference between the Most Significant Timestamp (MST) for prediction and the Least Significant Timestamp (LST) for prediction; $\tau = MST - LST$. MTS may also be regarded as the present time t_p: the time that separates the future and the past. We consider all records whose timestamp are greater than MST to have a weight of one ($f_\tau(a)$=1 for a<0). This case is under the assumption that all records in the future are the best predictors of the future. Our function f_τ is depicted in figure 1.

IV. Building a Predictive Model using Information Decay Technique

As Robert McDowell noted, *"Technology provides no benefits on its own; it is the application of technology to business opportunities that produces"*, the concepts describes so far provide no benefits on their own; it is the application of those concepts in real world (data mining) that reveal their benefits. For that reason, in this section, we expose the profit of using information decay approach on temporal data. Some researchers [2] discussed how to classify temporal data; however, they did not show how time-factor is accommodated in any applied classification technique and/or algorithm. We choose to apply our information decay concepts in decision tree based classifiers.

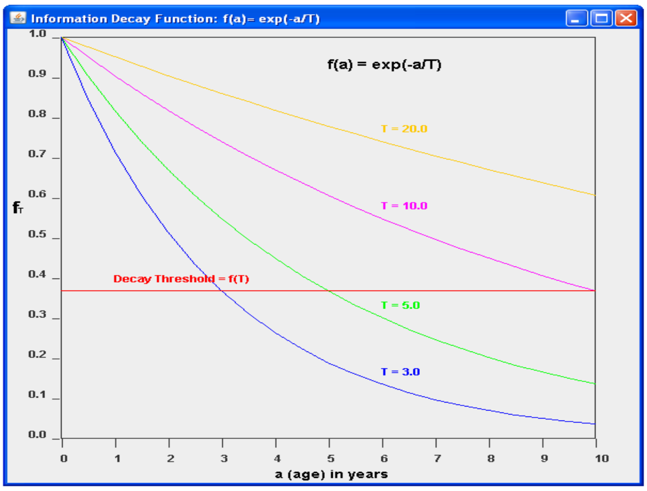

Figure 1. Information decay function $f_\tau(a)= exp(-a/\tau)$

A. Information Decay Applied in Decision Tree Algorithm

The basic idea of decision tree algorithms [9], [10] is classification of data by means of constructing a decision tree. Decision tree algorithms are based on information measures, which are concrete quantitative measurements of target class purity or impurity in a tree node. Let $p(c_i|x)$ denote the fraction of records belonging to class c_i at a given node x. If D is the dataset associated with node x, and D_i the records classified in class c_i at node x, $p(c_i|x)$ is also known as the probability of predicting a class c_i at node x, and is calculated as:

$$p(c_i \mid x) = \frac{|D_i|}{|D|} \qquad (1)$$

Widely used measures [8] include: Entropy, Gini, and Classification Error; their functions are respectively given as:

$$\text{Entropy(x)} = \sum_{i=0}^{n|1} p(c_i \mid x) \log_2 [p(c_i \mid x)] \qquad (2)$$

$$Gini(x) = 1| \sum_{i=0}^{n|1} [p(c_i \mid x)]^2 \qquad (3)$$

$$Classification_{Error(x)} = 1| \max[p(c_i \mid x)] \qquad (4)$$

It is obvious that $p(c_i|x)$ is the only variable parameter of the above mentioned information measures. The probability $p(c_i|x)$ is with certain possibility true if the used data recorded things, facts, or events that don't change with time; that is, the used data is static in nature. In reality, things change over time, and result in recording of temporal data. Therefore, for temporal data, this is not the exact probability that can be used for predicting the future; because, as time pass, the information that is being contributed by each record to predict future facts or events decays. The approximation of the exact probability can be found by decaying $p(c_i|x)$ using a decay function $f(a)$= $exp(-\lambda a)$, given the best approximation of the mean lifetime τ of data records. However, the lifetime of records is application domain dependent, and therefore, in this paper, its determination is reserved for future researches. In our case, we assume that the data miner has enough knowledge of the application domain to determine the lifetime of the used data. From equation (1) $|D_i|$ is the number of records belonging to class c_i, under the assumption that each record has a count of one, which means a record has 100% information on predicting what will happen in the future. In temporal data this information is decayed, and we therefore derive a new value of $p(c_i|x)$. Let m denote the number of records in D_i and $f_\tau(a_k)$ the decayed information of record R_k in D_i where a_k is the age of R_k. For temporal data with mean lifetime τ, $p(c_i|x)$ is computed as:

$$p(c_i \mid x) = \frac{\sum_{k=1}^{m} f_\tau(a_k)}{|D|} \tag{5}$$

In this case, historical data involved in building and verifying the model is not assumed to be the best estimator of what will happen in the future, rather, only its relevance in predicting the future is extracted and used. The benefits of this concept are revealed in the next subsection.

B. Experiment and Results

To demonstrate the advantages of information decay in mining temporal data, we needed to build to predictive models based on decision tree algorithm; one by information measure using equation (1), which is the existing approach, and the other using equation (5), which is the new approach presented in this paper. Using data available in the Sales History (SH) schema of the Oracle Database, we conducted a data mining experiment on predicting which of the customers are likely to increase spending if given an affinity card. In those data, a positive class (a target value of 1) is assigned to customers who increased spending with an affinity card, and a negative class (a target value of 0) is assigned to customers who did not increase spending. The experiment was conducted in three phases including: data preparation and understanding, models building, and models Evaluation, and they are explained bellow.

1. Data Preparation and Understanding

From the SH schema, demographic data about customers who have used an affinity card in the past were selected for 940 customers. The last time where a customer purchased goods was added to the data as the timestamp of when he/she did or did not increase spending. The dataset was found to be stamped from Mars 31, 1998 to June 30, 2000. The data were sorted in ascending order of their timestamp, and were divided into a training set (748 records) and a test set (192 records). A one-month gap (31-JAN-2000 to 29-FEB-2000) was used as a time separation between the two dataset. The first dataset that is stamped from 31-JAN-1998 to 31-JAN-2000, which is the build or training set, was assumed to be in the past, and the dataset that is stamped from 29-FEB-2000 to 30-JUN-2000, which is the test set, was assumed to be in the future. The one-month gap was assumed to be the time when the required models are being built and evaluated.

2. Models Building

Building a predictive model requires choosing a classification technique with a relavant data mining algorithm. Being the most successful machine learning technique in traditional AI in terms of practical application, the decision tree technique has been chosen along with its ID3 algorithm for our experiment. Using the training dataset, a Java implementation of the algorithm was iteratively used to create different models by changing the tree-settings parameters, and however, ignoring the timestamp attribute; i.e. equation (1) was used in information measures. Resulting models were tested using the test dataset, and were compared based on their accuracy, and the best performing model, along with its parameters, was selected and named Model A (M_A). The selected tree-settings parameters were used to build another model; however, this time, the timestamp attribute was enabled and equation (5) was used in information measures. We greatly benefited from the *java.util.Date* class, which implements date and time in milliseconds; it provided a great benefit in implementing the information decay function by changing each timestamp in milliseconds in order to easy the required calculations. A three years period was estimated to be the mean lifetime of records and it was changed in milliseconds to normalize its value. Under the assumption that the model is being built in the present for predicting the future, 1st February, 2000 was used as the reference to the present (the time when the model is being built.) The resulting model was named Model B (M_B).

3. Models Evaluation

Using the training set, Model A and Model B were compared with the usage of three types of popular test metrics for classification models including: prediction accuracy, F_1 measure, and receiver operating characteristics (ROC). The training set consisted of 45 positive cases, which have a target value of 1 (customers who increased spending) and 147 negative cases, which have a target value of 0 (customers who did not increase spending). Due to the class imbalance problem in the build dataset a cost-sensitive testing was applied to find the highest F_1 measure value (a high value of F1-measure ensures that both precision and recall are reasonably high.) The cost of predicting a False Positive (FP) was fixed to 1, and that of predicting a False Negative (FN) was incremented from 1 to 4 with and increment value of 0.5, and for each FN to FP ratio the prediction accuracy (ACC) and F_1-measure (F_1) values were calculated; the results are shown in table 2. As a high value of F_1-measure ensures that both precision and recall are reasonably high, the ratio of FN to FP that gives the highest F_1-measure values for both the models was selected, and the receiver operating characteristic (roc) curves of the two classifiers were plotted as shown in figure 2.

Table 2. Cost-sensitive testing of Model A and Model B

	FN/FP	1.0	1.5	2.0	2.5	3.00	3.5	4.0
M_A	ACC	0.77	0.77	0.71	0.74	0.71	0.71	0.69
	F_1	0.50	0.55	0.49	**0.57**	0.55	0.55	0.56
M_B	ACC	0.76	0.74	0.76	0.76	070	0.76	0.66
	F_1	0.45	0.55	0.61	**0.61**	0.56	0.50	0.54

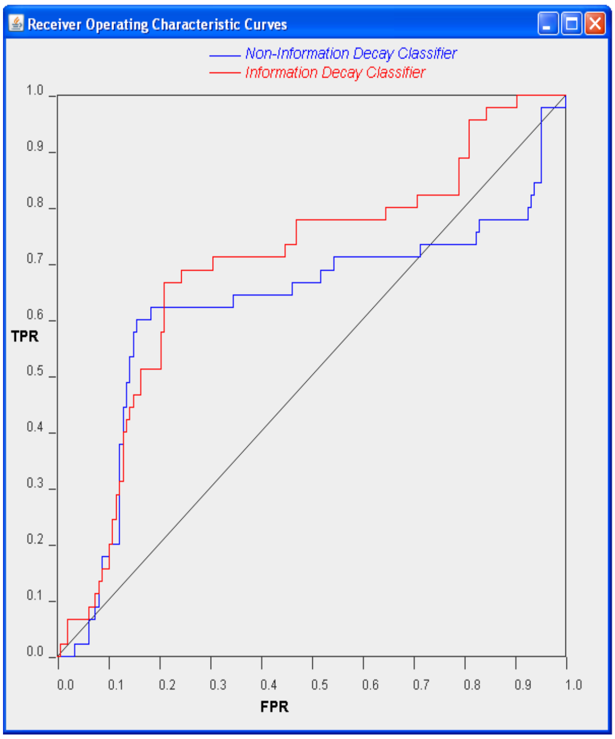

Figure 2. Receiver Operating Characteristic Curves

V. RESULTS DISCUSSION

Our experimental results indicate that for high values of F_1-measure Model A performs better than Model B, which means that Model A is preferable over Model B as it is the one that generates the higher values for recall and precision. In figure 2, it is obvious that the receiver operating characteristic curves of the information decay classifier generates an area under the curve that is bigger that the area under the curve of the noninformation decay classifier. Therefore, using information decay approach, in predictive data mining on temporal data, results in creating many models that have higher performance than using noninformation decay technique. The information decay technique may result in improving the performance of the algorithm as records that have a decay weight that is lesser than a given weight threshold (for instance, is a threshold used 37% to ignore records whose age is greater than the mean lifetime) are eliminated automatically, therefore, allowing the computation to be performed only on those records that are relevant to the prediction of the future events or facts. In generating rules, Model B generated rules starting from those attributes that mostly characterized recent facts or event and resulted in generating fewer rules than explain the change in behaviors of the system, hence, inferring the relationships of contextual and temporal proximity, which indicate cause-effect association. However, Model A generated rules that represents static nature of the data.

Therefore, using temporal data, it is better to use information decay classification technique in building predictive models, as it results in models that provide a dynamic view of an evolving business domain, rather than providing a static view of the application domain.

I. CONCLUSION

The growth of research and applications of temporal database has motivated an urgent need for techniques to deal with the problem of knowledge discovery from temporal databases. It is in that regards that, in this paper, we poured our efforts on searching a new practical approach to mine temporal data. The information decay technique was therefore identified as our new data mining approach to predict the future using temporal data. We have discussed on how to apply the information decay function in a decision tree classification technique and its benefits during the model building. It has been shown that the new technique produces higher values for the recall and precision using F_1 measure. Based on the observed benefits of the information decay mining technique, we encourage other researchers to provide practical ways of mining temporal data using the information decay function in other classification techniques such as artificial neural networks, Bayesian classifiers, etc.

REFERENCES

[1] X. Chen, and I. Petrounias, “A Framework for Temporal Data Mining,” Database and Expert Systems Applications, Vol. 1460, pp.796–806 , 1998.

[2] F.Roddick, and M. Spiliopoulou, “A Survey of Temporal Knowledge Discovery Paradigms and Methods,” IEEE Transactions on Knowledge and Data Engineering, Vol.14, No.4, pp.750–767, July/August 2002.

[3] D. Pan, and Y. Pan , “A Genaral Framework on Temporal Data Mining,” Proceeding of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

[4] G. Zeira, O. Maimon, M. Last, and L. Rokach, “Change Detection in Classification Models Indeced from Time Series Data,” Data Mining in Time Series Database, Marchine Perception in Artificial Intelligence, Vol.57, pp.101–125

[5] J. Debenham, “Knowledge Decay in a Normalised Knowledge Base,” Database and Expert Systems Applications, Vol. 1873, pp. 417–426, 2000.

[6] G. Cormode, V. Shkapenyuk, D. Srivastava, and B. Xu, “Forward Decay: A Practivacl Time Decay Model for Streaming Systems,” Proc. ICDE’09, IEEE 25th International Conference on Data Engineering, March 29 2009-April 2 2009, pp. 138–149 , doi: 10.1109/ICDE.2009.65.

[7] F. Hornick, E. Marcade, and S. Venkayala, “Java Data Mining, Strategy, Standard, and Practive,” Morgan Kaufmann, 2007, pp. 150–157.

[8] P. Ning Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining , Pearson Education, 2006, pp. 23–24, pp.135–141

[9] T. Munakata, Fundamentals of the New Artificial Intelligence, Second Edition, Springer, 2008, pp.195–2202.

[10] M. Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms,” John Wiley & Sons, 2003.

Lasheng Yu, Vice professor in Central South University of China, ACM and CCF member, ACM/ICPC golden medal coach. He received the B.Sc. degree in Computer Science, the Master degree and a Ph.D. degree in Control Theory and Control Engineering from Central South University. He is the editor of Journal of Convergence Information Technology and Advances in Information Sciences and Service Sciences etc. He has published at least 50 papers on Agent technologies or Algorithms, and has published 3 books. He has organized and implemented many projects which have created great achievements in the society. His main research interests include agent technologies and applications, structure and algorithm, social computing etc.

Mukwende Placide, Lecturer at Adventist University of Central Africa. He got his Bachelor of Technology in Information Technology at Vellore Institute of Technology, INDIA, and got his Master degree in Computer Science and Technology in Central South University, His main research interests include agent technologies and applications, data mining.