International Journal of Multidisciplinary and Current Research

Research Article

ISSN: 2321-3124
Available at: http://ijmcr.com

Finding Significant Frequency Changes in Large Databases

D.Veerabhadra Babu[1],Dr.K.Fayaz [2] and D. William Albert[1]

[1]Department of Computer Science & IT, Mahatma Gandhi University, Meghalaya, India
[2]Department of Computer Science & Technology, SK University, Anantapuram, India

Accepted 15 Aug 2014, Available online 25 Aug 2014, **Vol.2** (**July/Aug 2014 issue)**

Abstract

Data mining can discover significant frequency changes between two large databases. When multiple data sources are used it is important sometimes to know the subtle differences between the databases. Such knowledge can help in making well informed decisions. This kind of mining is known as contrast mining. Pattern which is not in one dataset and present in another dataset is known as jumping emerging pattern. On the other hand, the pattern which has less frequency in first dataset and more frequency in other dataset is known as emerging pattern. In this paper, a general framework for building robust and accurate classifier is presented which guides in making EP based classifier. Afterwards, an algorithm is proposed for making a model that can help in classification of testing instances. We have built a prototype application that demonstrates the proof of concept. The empirical results are compared with other classifiers like NB, SVM and C4.5.

Keywords: *Data Mining, emerging patterns, SVM, NB*

1. Introduction

Data mining has become ubiquitous in enterprises for discovering knowledge and making expert decisions. It is very important in the areas like finance, banking, and insurance where monetary transactions are involved. In the process of extracting actionable knowledge, sometimes, it is imperative to contrast facts obtained from one dataset with another dataset. The facts thus known can be related to various time periods, different geographical locations, and across different classes. The contrast pattern mining is useful in finding rare classes, discovering the inadequacy of data, finding abundance, providing ranking besides combination of them. Contrast in this context refers to a classification rule or discriminator or difference or change. Therefore contrast data mining can be used to detect changes, extract class based association rules, extract high confidence patterns or emerging patterns or discriminative patterns or top k patterns. Contrast data mining is also related to concept drifting [1].

Emerging patterns are the patterns with different frequency in different databases. Considering D1 and D2 are two datasets, an emerging pattern X is the item set that exhibits large growth rate from D1 to D2. The growth rate is considered 0 if support(X) is 0 in D1 and D2. The growth rate is said to be infinite if support(X) is 0 in one dataset and support(X)>0 in other dataset. Otherwise the growth rate is computed as support(X) in D2/support(X) in D1. When a pattern is present in dataset and absent in another dataset, it is said to have infinite growth rate. Such emerging pattern is known as Jumping Emerging Pattern (JEP). The emerging patterns can reveal strong contrast knowledge between databases. Our contributions in this paper are as described here.

1. We have proposed an algorithm for building a robust classifier which is based on emerging patterns. The algorithm is named as Emerging Pattern Based Classifier (EPBC). The EPBC classifier is used to classify test instances. A class with highest score for given testing instance is associated with that instance.
2. We have built a prototype application that demonstrates the usefulness of the proposed algorithm. The prototype takes support from Weka tool for basic operations and visualization.

The remainder of the paper is structured as follows. Section II reviews literature on contrast mining. Section III provides the proposed framework for finding frequency changes in large databases. Section IV presents experimental results while section V concludes the paper.

2. Related Works

This section review relevant literature. There are many concepts pertaining to contrast pattern mining. Contrast sets [6], discriminative rules [5], [4], disjunctive version spaces [3] and version spaces [2] are examples for emerging patterns which are nothing but contrast patterns. These patterns can provide strong contrast knowledge between databases. Contrast pattern mining is best used to build classifiers. There are many classification models found in the recent literature. They include nearest neighbor classifiers [13], instance-based learning algorithms [12], statistical methods and log-linear models [11], Bayesian methods [10], genetic algorithms [9] and neural networks [8], [7]. In addition to these classification trees and decision trees are found in the literature [17], [16], [15] and [14].

Emerging patterns are the patterns that show significant changes between two databases. These patterns are very useful to build classification models. Emerging patterns can be used to build EP-based classifiers for classification purposes. Such classifiers should have strong contrast knowledge between two databases. They are useful in making well informed decisions. In the same fashion decision trees, neural networks, SVM can also be used for classification.

3. Discovering Frequency Changes in Large Databases through Contrast Mining

This section provides details about the discovery of frequency changes in large databases through contrast mining.

A. General Framework for Building EP-Based Classifier

This section describes the proposed framework for building classifier that can be used to discover patterns in given two datasets. The knowledge thus discovered represents the significant frequency changes in large databases. The framework is illustrated as shown in Figure 1.

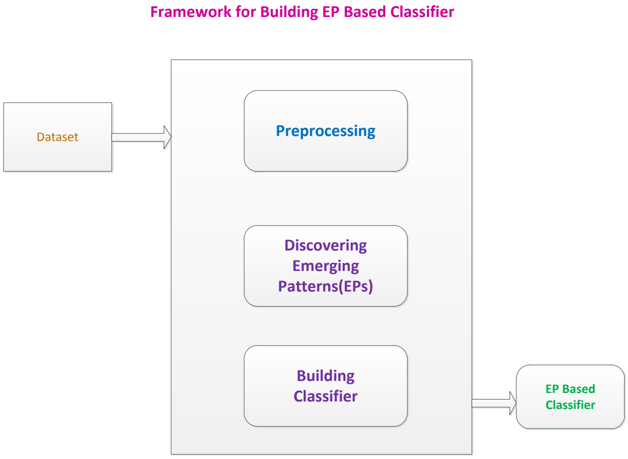
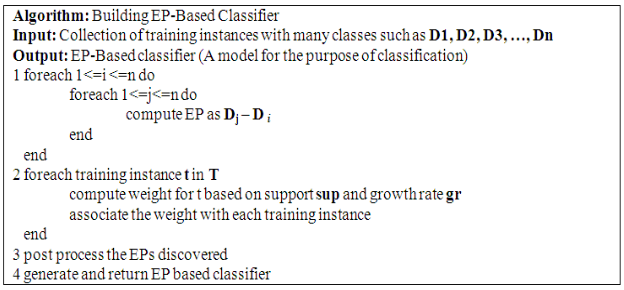

Framework for Building EP Based Classifier
Dataset
Preprocessing
Discovering Emerging Patterns(EPs)
Building Classifier
EP Based Classifier

Figure 1 – Illustrates the general procedure for building EP based classifier

EP based classifier is based on the strong contrast knowledge between two datasets. Therefore it has better utility when compared with traditional classifiers like NB, C4.5, and SVM. The proposed framework for building classifier has a pre-processing step. It is meant for taking training dataset as input and converting continuous attributes into discrete attributes. Then EP mining algorithm is applied on the discretized data in order to EPs for each class of data. Support and growth rates are described in the previous section. These statistical measures are used to associate weights with tuples present in training dataset. Thus a weighting model is used in building EP based classifier. The classifier thus built is used further for classifying emerging patterns. The classifier classifies all test instances present in test data.

B. Algorithms for Building Classifier and Extracting EPs

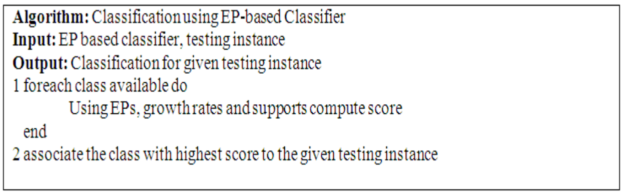
We built two algorithms that are meant for building classifier and classifying EPs respectively. The first algorithm takes a set of training instances as input and generates an EP based model that can be used to classify test instances based on EPs. Figure 2 presents the pseudo code for building EP based classifier.



```
Algorithm: Building EP-Based Classifier
Input: Collection of training instances with many classes such as D1, D2, D3, …, Dn
Output: EP-Based classifier (A model for the purpose of classification)
1 foreach 1<=i<=n do
        foreach 1<=j<=n do
                compute EP as $D_j - D_i$
        end
  end
2 foreach training instance t in T
        compute weight for t based on support sup and growth rate gr
        associate the weight with each training instance
  end
3 post process the EPs discovered
4 generate and return EP based classifier
```

Figure 2 – Illustrates algorithm for building EP-based classifier

The proposed algorithm takes a dataset with many classes of data. For each class of data EPs are computed. With the help of statistical analysis based on support and growth rate values weights are associated with each training instance. Then the EPs that have been discovered are subjected to post process. Finally a model is built ready to serve classification purposes. The output of the algorithm is the classifier that can classify when test data is given to the application. Having built a robust classifier, now it is the time to use this classifier to classify testing set. Figure 3 shows the algorithm used to classify testing dataset based on the classifier that has been built using the algorithm presented in Figure 2.



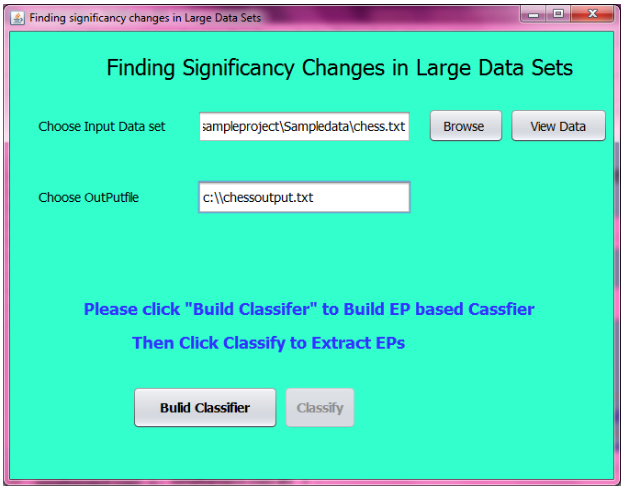
```
Algorithm: Classification using EP-based Classifier
Input: EP based classifier, testing instance
Output: Classification for given testing instance
1 foreach class available do
        Using EPs, growth rates and supports compute score
  end
2 associate the class with highest score to the given testing instance
```

Figure 3 – Illustrates algorithm for classification using EP-based classifier

The algorithm presented in Figure 3 takes EP based classifier which has been built using the algorithm shown in Figure 2. Along with it, this algorithm also takes a testing instance. As the classifier has been trained with trained data the algorithm can make use of growth rates, supports, and EPs in order to compute score for each and every class present. After completion of this iterative process for each class, the algorithm will come to know the class with highest score for given testing instance. Then it associates that class with the testing instance. This concept is known as classification. The EP based classification has proved to be effective when compared with traditional classifiers as it gets knowledge pertaining to strong contrast information discovered in the form of EPs.

4. Prototype Application

A prototype application is built to demonstrate the proof of concept of the proposed generalized framework for building robust classifier based on EPs. The application is built using Java programming language. The application is user friendly with visualization of performance of various algorithms. Swing API of Java are used to build the user interface while JBDC API is used interact with dataset.


Finding significancy changes in Large Data Sets

Finding Significancy Changes in Large Data Sets

Choose Input Data set: sampleproject\Sampledata\chess.txt

Choose OutPutfile: c:\\chessoutput.txt

Please click "Build Classifer" to Build EP based Cassfier

Then Click Classify to Extract EPs

Figure 4 – Shows UI for building EP based classifier and then extract EPs

The prototype application is used to perform experiments that bestow emerging patterns. The performance of the EP based classifier is compared with other existing classifiers. The ensuring section shows the experimental results.

5. Experimental Results

This section presents experimental results that are obtained by running prototype application which supports building robust and accurate classifier which is EP based besides comparing the results with other traditional classifiers. Experiments are made by introducing noise and without noise on datasets. The following subsections describe more details on empirical results. Various datasets are taken from UCI machine learning repository [19]. The empirical results made with our proposed algorithm are compared with other classification algorithms such as NB, C4.5 and SVM. The proposed algorithm is EP based while other algorithms do not make use of any contrast patterns. Figure 5 visualizes the accuracy comparison of the EP based algorithm with traditional algorithms.

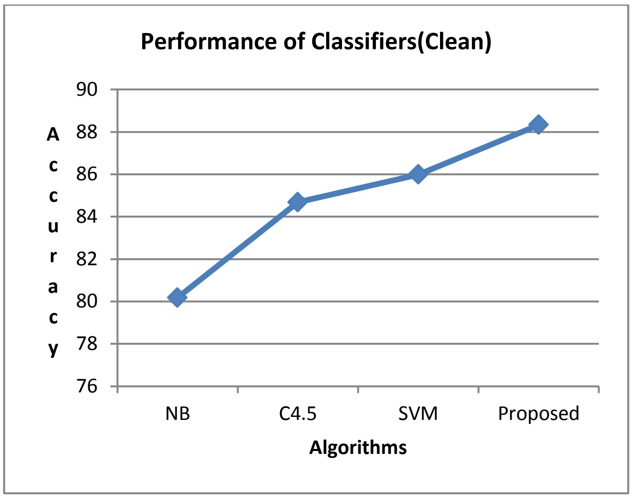


Figure 5 – Performance of classifiers (no noise introduced)

As shown in Figure 5 it is evident that the experimental results on datasets without noise (clean) are presented. The proposed algorithm which is EP based classifier shows clearly better performance over all traditional classifiers. The reason behind this is that EPs represent strong contrast knowledge of datasets which make the classifier robust. The algorithms such as NB, C4.5, SVM and the proposed EP based algorithm have performance in increasing order with respect to accuracy.

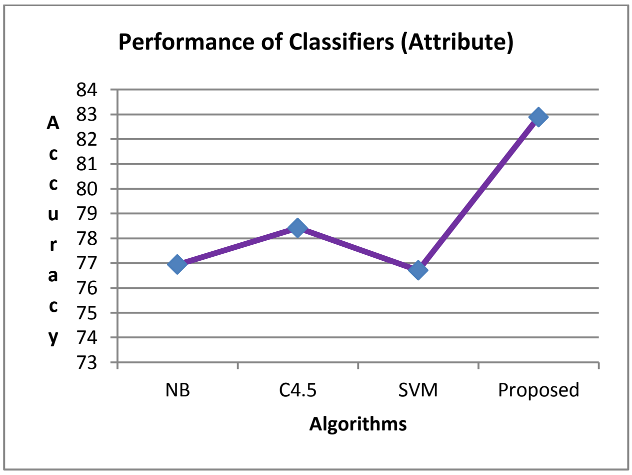


Figure 6 – Performance of classifiers (40% attribute noise)

As shown in Figure 6 it is evident that the experimental results on datasets with 40% noise on attributes are presented. The proposed algorithm which is EP based classifier shows clearly better performance over all traditional classifiers. The reason behind this is that EPs represent strong contrast knowledge of datasets which make the classifier robust to noise. The algorithms SVM, NB, C4.5 and proposed EP-based exhibit accuracy in increasing order respectively. Therefore it is evident that EP based classifier has edge over all other algorithms. In this case, SVM exhibits least performance as it is not robust to attribute noise.

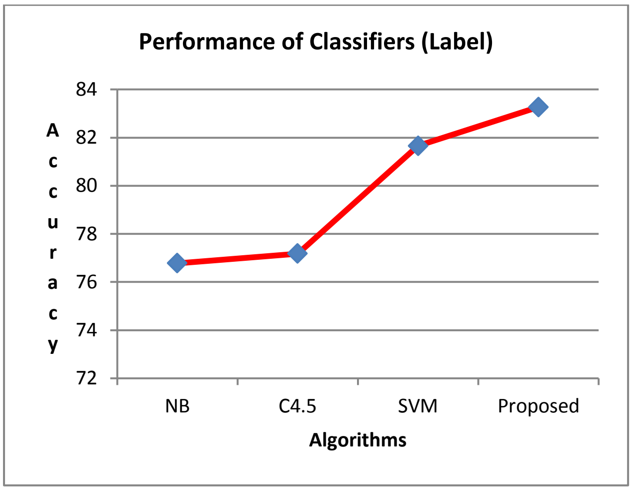


Figure 7 – Performance of classifiers (40% label noise)

As shown in Figure 7 it is evident that the experimental results on datasets with 40% noise on labels are presented. The proposed algorithm which is EP based classifier shows clearly better performance over all traditional classifiers. The reason behind this is that EPs represent strong contrast knowledge of datasets which make the classifier robust to noise. The algorithms NB, SVM, C4.5 and proposed EP-based exhibit accuracy in increasing order respectively. Therefore it is evident that EP based classifier has edge over all other algorithms. In this case, NB exhibits least performance as it is not robust to label noise.

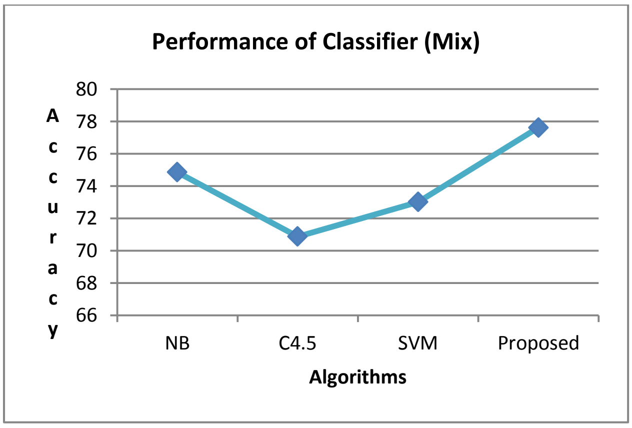


Figure 8 – Performance of classifiers (40% attribute and label noise)

As shown in Figure 8 it is evident that the experimental results on datasets with 40% noise on both attributes and labels are presented. The proposed algorithm which is EP based classifier shows clearly better performance over all traditional classifiers. The reason behind this is that EPs represent strong contrast knowledge of datasets which make the classifier robust to noise. The algorithms C4.5, SVM, NB and proposed EP-based exhibit accuracy in increasing order respectively. Therefore it is evident that EP based classifier has edge over all other algorithms. In this case, C4.5 exhibits least performance as it is not robust to mixed noise introduced to attributes and labels.

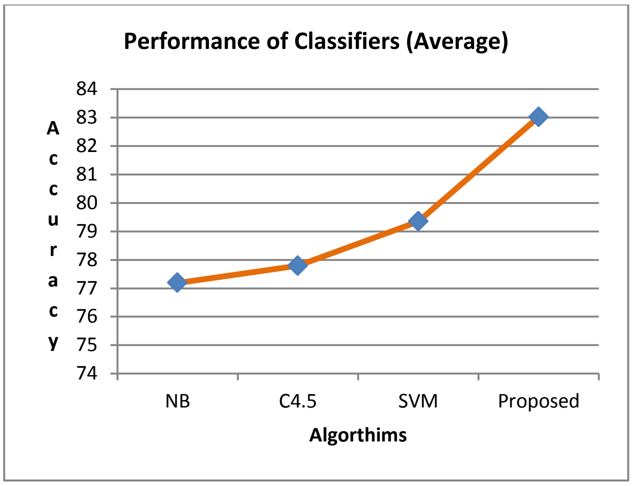


Figure 9 – Average performance of classifiers

As shown in Figure 9 it is evident that the experimental results on datasets with average accuracy. The proposed algorithm which is EP based classifier shows clearly better performance over all traditional classifiers. The reason behind this is that EPs represent strong contrast knowledge of datasets which make the classifier robust to noise. The algorithms NB, C4.5, SVM and proposed EP-based exhibit accuracy in increasing order respectively. Therefore it is evident that EP based classifier has edge over all other algorithms. In this case, NB exhibits least performance as it is not robust to noise introduced to attributes and labels.

Conclusions and Future Work

In this paper we have studied contrast pattern mining. Especially emerging patterns concept is used to build a robust classifier. The proposed classifier algorithm is known as EPBC (Emerging Pattern Based Classifier) which makes use of EPs for building a classifier. Then the classifier is used for classification of patterns. Such patterns hold strong contrasting knowledge which is very useful. Four kinds of datasets are used. Datasets with no noise introduced, datasets with 40% noise introduced to attributes, datasets with 40% introduced to labels, datasets with 40% noise introduced to both attributes and labels. The prototype application is used to test the efficiency of the proposed algorithm. As the results revealed, the proposed EP based classifier exhibited higher performance when compared with traditional classifiers like NB, C4.5 and SVM. When no noise is introduced NB, C4.5, SVM and proposed classifier exhibited performance in the increasing order respectively. When 40% attribute noise is introduced the increasing performance order is SVM, NB, C4.5 and proposed algorithm. From this it can be concluded that NB is not robust to the attribute noise. When 40% label noise is introduced the increasing performance order is NB, C4.5, SVM and proposed algorithm. When 40% attribute and label noise is introduced together, the increasing performance order is C4.5, SVM, NB and proposed algorithm. From the results it can be understood that the proposed EP based algorithm outperformed the other classifiers like NB, C4.5 and SVM. This is because the EPs provide strong knowledge base pertaining to emerging contrast patterns between datasets. With all experiments, the proposed EP based algorithm has shown high performance for this reason. The experiments without noise introduced reveal that NB has least performance. When 40% attribute noise is introduced SVM exhibited least performance. When 40% label noise is introduced NB exhibited least performance. When 40% attribute and label noise is introduced C4.5 exhibited least performance. Our future work is to explore discovering contrast mining in a single step without the need for extracting frequent patterns.

References

[1] Michael A. Osborne.. (2010). LEARNING FROM DATA STREAMS WITH CONCEPT DRIFT. ICML. 2 (5), 1-160.

[2] Tom Mitchell. Generalization as search. *Artificial Intelligence*, 18(2), 1982.

[3] M. Sebag. Delaying the choice of bias: A disjunctive version space approach. In *Proc. 13th Int'l Conf. on Machine Learning*, pages 444–452. Morgan Kaufmann, 1996.

[4] Jiawei Han, Yandong Cai, and Nick Cercone. Knowledge discovery in databases: An attribute-oriented approach. In Li-Yan Yuan, editor, *Proc. 18th Int'l Conf. on Very Large Databases*, pages 547–559, San Francisco, U.S.A., 1992. Morgan Kaufmann Publishers.

[5] Jiawei Han, Yandong Cai, and Nick Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5(1):29–40, 1993.

[6] Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[7] Christopher M. Bishop and Chris Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[8] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[9] A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Spinger-Verlag, Berlin, 2002.

[10] Peter Cheeseman and John Stutz. Bayesian classification (autoclass): Theory and results. In *Proc. 2nd Int'l Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pages 153–180. AAAI/MIT Press, 1996.

[11] Ronald Christensen. *Log-Linear Models and Logistic Regression*. Springer, 1997.

[12] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[13] Belur V. Dasarathy. *Nearest neighbor norms: NN pattern classification techniques*. IEEE Computer Society Press, Los Alamitos, CA, 1991.

[14] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[15] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[16] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[17] RuleQuest. See5/c5.0, 2000. RULEQUEST RESEARCH data mining tools [http://www.rulequest.com/].

[18] Patrick Laube et al.. (2009). Spatial support and spatial confidence for spatial association rules. Australian Reserach Council(ARC). 2 (5), 1-21.

[19] Bache, K. & Lichman, M. (2013). Machine Learning Repository. Available: https://archive.ics.uci.edu/ ml/datasets.html. Last accessed 5th June 2014.

Authors

D.Veerabhadra Babu Post graduated in Master of Computer Applications in 2005 and Master of Technology in Information Technology in 2011 from K.S.O.University, Karnataka and is a Research Scholar in Mahatma Gandhi University, Meghalaya in Data Mining specialization. Nearly 15 National & International technical papers published and guided M.Tech(8), M.C.A(40). and B.Tech(35) Students in their Project works The area of interests includes Data Warehousing & Mining, Software Engineering, Software Testing Methodologies, Software Project Management, Human Computer Interface, Management Information Systems, etc.

Dr.K.Fayaz is working as a faculty of system (System In-charge) in Sri Krishnadevaraya Institute of Management (SKIM), Sri Krishnadevaraya University, Anantapuramu. He acquired B.Sc., from S.K.University, Post Graduate Diploma in Computer Application from JNTU, Anantapuramu and Master of Information Technology from Manipal Academy of Higher Education (MAHE) Deemed University. Manipal. Subsequently has done M.Phil and Ph.D from S.K.University in Computer Science & Technology. He has published more than 10 papers in Computer Science International Journals. He has attended and presented papers in several International, National Conferences, Seminars, Workshops.

D.William Albert, is a Research Scholar in Mahatma Gandhi University, Meghalaya in Data Mining specialization. Basically a B.Sc. Science Graduate and did his Post-graduation Masters in Computer Science in 2004 and Master of Technology in Computer Science & Engineering in 2006. He is a Life Member of profession body such as **I**ndian **S**ociety for **T**echnical **E**ducation, New Delhi. More than 15 National & International technical papers published and guided M.Tech, M.C.A. and B.Tech Students in their Project works. The area of interests includes Data Warehousing & Mining, Software Engineering, Software Testing Methodologies, Software Project Management, Distributed Operating System, Human Computer Interface, E-commerce, Enterprise Resource Planning, etc.