APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE

A. B. Devale[1] and Dr. R. V. Kulkarni[2]

[1]Arts, Commerce, Science College, Palus Dist. Sangli, Maharashtra
amol.devale99@gmail.com
[2]Shahu Institute of Business Research, Kolhapur, Maharashtra
drrvkulkarni@siberindia.co.in

ABSTRACT

Knowledge discovery in financial organization have been built and operated mainly to support decision making using knowledge as strategic factor. In this paper, we investigate the use of various data mining techniques for knowledge discovery in insurance business. Existing software are inefficient in showing such data characteristics. We introduce different exhibits for discovering knowledge in the form of association rules, clustering, classification and correlation suitable for data characteristics. Proposed data mining techniques, the decision- maker can define the expansion of insurance activities to empower the different forces in existing life insurance sector.

KEYWORDS

Insurance, Association rules, Clustering, Classification, Correlation, Data mining.

1. INTRODUCTION

Data mining can be defined as the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns. In the insurance industry, data mining can help firms gain business advantage. For example, by applying data mining techniques, companies can fully exploit data about customers' buying patterns and behavior – as well as gaining a greater understanding of their business to help reduce fraud, improve underwriting and enhance risk management. This paper discusses how insurance companies can benefit by using modern data mining methodologies and thereby reduce costs, increase profits, acquire new customers, retain current customers and develop new products. Data mining methodology often can improve upon traditional statistical approaches to solving business solutions. For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. Data mining often can improve existing models by finding additional, important variables, identifying interaction terms and detecting nonlinear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. Specifically, data mining can help insurance firms in business practices such as:

- Acquiring new customers.
- Retaining existing customers.
- Performing sophisticated classification
- Correlation between Policy designing and policy selection

DOI : 10.5121/ijdkp.2012.2404

1.1 Acquiring New Customers

An important business problem is the acquisition of new customers. Although traditional approaches involve attempts to increase the customer base by simply expanding the efforts of the sales department, sales efforts that are guided by more quantitative data mining approaches can lead to more focused and more successful results. A traditional sales approach is to increase the number of policyholders by simply targeting those who meet certain policy constraints. A drawback to this approach is that much of the marketing effort may yield little return. At some point, sales become more difficult and greater marketing budgets lead to lower and lower returns. Hence in this situation it is important to identify population segments among already insured customers through which uninsured customers could be targeted. A statistical technique called "cluster analysis," sometimes used in the private sector to identify various market segments, was used to identify target groups of uninsured adults based on the previous available data of policy holders. Clustering is a technique of partitioning or segmenting the data into groups that might or might not be disjointed. The clustering usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters. Since clusters are not predefined, a domain expert is often required to interpret the meaning of the created clusters.

Example 1.1

Insurance companies can create special catalogs targeted to various demographic groups based on attributes such as income, occupation and age as physical characteristic of potential customers. The company then can perform a clustering of potential customers based on determined attribute values to create new catalogs. The results of the clustering exercise can be then used by management to create special catalogs for different policies and distribute them to the correct target population based on the cluster for that policy.

An insurance company can group its customers based on common features. Company management does not have any predefined for this label. Based on the outcome of the grouping they will target marketing and advertising campaigns to the different groups for a particular type of policy.

Table 1.1 Sample data for example

Age	Occupation	Income	Education
35	Employee	15,000/-	Graduate
25	Employee	10,000/-	Graduate
55	Employee	65,000/-	Post-Graduate
45	Employee	45,000//	Post-Graduate
40	Business	70,000/-	Matriculate
35	Business	90,000/-	Graduate

The information they have about the customers include Age, Occupation, Income and education. Depending on the type of policy, not all attributes are important. For example, suppose advertising only for policy of Life security, we could target the customers having less income and occupation as employee. Hence the first group of people, is of younger employees having college degree, is suitable for Life security policies. The second group has higher qualification and also higher income is suitable for tax benefit policies, while last group has businessmen with higher income but low qualification and is suitable for investment policies.

> DEFINITION 1.1. Given a database D = {t1,t2,…,tn } of tuples and an integer value k, the clustering problem is to define a mapping f : D → {1,…,k}where each ti is assigned to one cluster kj, 1≤ j ≤ k.A cluster kj, contains precisely those tuples mapped to it that is,kj = { ti | f(ti) = kj, 1≤ i ≤ n, and ti Є D}

Algorithm 1.1 k-means Clustering

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached

```
Input:
        D= {t1, t2, t3,...,tn}    //Set of elements
        k   //Number of desired clusters
Output:
        K   //set of clusters
Algorithm:
        assign initial values for means m1, m2, …, mk;
        repeat
assign each item ti to the cluster which has closest mean;
calculate new mean for each cluster;
        until convergence criteria is met.
```

Note that the initial values for means are arbitrarily assigned and the algorithm could stop when no or very small number of tuples are assigned to different clusters. As per the algorithm, first we have to find mean of each cluster. Hence accordingly mean for first cluster is 30, Employee, 13000, Graduate in terms of Age, Occupation, Income and Education. Similarly mean for second cluster is 50, Employee, 50000, Post-Graduate while the same for third cluster is 37, Business, 80000, Graduate. Suppose a customer with age 36, occupation Employee, Income 14000 and education Graduate will provide differences 6,0,1000,0 with average difference of 252 for first cluster. Similarly it provides average difference of 9003 and that of 16001 for third cluster. Hence observing the above means, it is clear that the closest cluster for this customer is the first cluster i.e. of life security policy. Once the customer is added to one of the clusters its new mean will be automatically calculated.

1.2 Retaining Existing Customers

As acquisition costs increase, insurance companies are beginning to place a greater emphasis on customer retention programs. Experience shows that a customer holding two policies with the same company is much more likely to renew than is a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding less than three. By offering quantity discounts and selling bundled packages to customers, such as home and auto policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

So we have determined the frequent item sets based on a predefined support. We have all the riders that are often sold together. We need to find all the associations where customers who bought a subset of a frequent item set, most of the time also bought the remaining items in the same frequent item set. Association refers to the data mining task of uncovering relationships among data. Data association can be identified through an association rule

Example 1.2

Insurance companies can use association rules in market basket analysis. Here the data analyzed consist of information about what policies customer purchases. The insurance company can generate association rules that show what different policies are purchased with a specific policy. Based on these facts, company tries to capitalize on the association between different policies that are sold for different purposes. Experience shows that a customer holding two policies with the same company is much more likely to renew than is a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding less than three. By offering quantity discounts and selling bundled packages to customers, such as life security and investment policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

Table 1.2 Sample data for example

Transaction	Items
T1	Life security, Market based
T2	Market based
T3	Investment
T4	Market based, Tax Benefit, Investment
T5	Market based, Tax Benefit
T6	Market based, Tax Benefit
T7	Life security, Market based, Tax Benefit, Investment
T8	Life security, Tax Benefit
T9	Life security, Market based, Tax Benefit
T10	Life security, Market based, Tax Benefit

A database in which an association rule is to be found is viewed as a set of tuples, where each tuple contains a set of items. Here there are ten transactions and four items: {Life security, Market based, Tax Benefit, Investment} which are to be considered as {S1,S2,S3,S4}.

Now we need to find all the situations where customers who bought a subset of a frequent itemset, most of the time also bought the remaining items in the same frequent itemset. Given a frequent itemset, say (S1, S2, S3), if a customer who buys a subsert formed by S1 and S2, also buys S3 80% of the times, then it is worth to consider the rule. This percentage is called the confidence of the rule and is defined as the ratio of the number of transactions that include all items in a particular frequent itemset to the number of transactions that include all items in the subset.

Let's consider the same insurance example below. We want to find the association rules that meet the following requirements:

Support - 30% - Only the riders that are bought together by at least 3 customers are considered.
Confidence - 90% - The association rule has to be true in 90% of the transactions

Case1: (S1, S3) → (S2)(S1, S3) was bought by 5 customers but only 3 of them also bought S2. Confidence is 60%.

Case2: (S1, S2) → (S3)(S1, S2) was bought by 3 customers and all 3 of them bought S3 as well. Confidence is 100%. So this rule has a very strong confidence (above 90%) and has to be considered.

> DEFINITION 1.2.1. Given a set of items I = {I1,I2,…,Im } and a database of transactions D = {t1,t2,…,tn } where ti = {Ii1,Ii2,…,Iik } and Ijk Є I, an association rule is an implication of the form X ⇨ Y where X,Y (I are sets of items called itemsets and X ∩ Y = ø
>
> DEFINITION 1.2.2. The support (s) for an association rule X ⇨ Y is the percentage of transactions in the database that contain X U Y.
>
> DEFINITION 1.2.3. The confidence or strength (α) for an association rule X ⇨ Y is the ratio of the number of transactions that contain X U Y to the number of transactions that contain X.
>
> DEFINITION 1.2.4. Given a set of items I = {I1,I2,…,Im } and a database of transactions D = {t1,t2,…,tn } where ti = {Ii1,Ii2,…,Iik } and Ijk Є I, an association rule is to identify all association rules X ⇨ Y with a minimum support and confidence. These values (s, α) are given as input to the problem.

Algorithm 1.2Apriori Algorithm

The Apriori algorithm is the most well known association rule algorithm and is used in most commercial products.

```
 Input:
        Li-1    //Large itemsets of size i - 1
Output:
        Ci   //candidates of size i
Algorithm:
        Ci = ø;
        for each I Є Li-1 do
           for each J ≠ Є Li-1  do
if i – 2 of the elements in I and J are equal then
    Ck = Ck U {I U J};
```

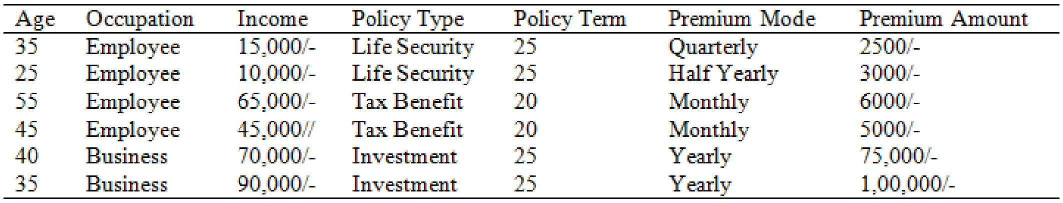
1.3 Classification: Segment Databases

To improve predictive accuracy, databases can be segmented into more homogeneous groups. Then the data of each group can be explored, analyzed and modeled. Depending on the business question, segmentation can be done using variables associated with risk factors, profits or behaviors. Segments based on these types of variables often provide sharp contrasts, which can be interpreted more easily. Classification maps data into predefined groups or segments. Classification algorithms require that the classes be defined based on data attributes values. They often describe these classes by looking at the characteristics of data already known to belong to the classes. As a result, insurance companies can more accurately predict the likelihood of a policy based on the premium mode, premium amount, policy period depending upon age, income and occupation.

Example 1.3

Insurance company can find a segment based on the income, preferred premium mode and premium amount. Such patterns can be stored in database. So while selling a specific policy to customer, agent can get the information of customer like income and age. This pattern can be compared to entries in a database and agent can suggest premium modes, premium amount and policy period to customer based on matched patterns.

Table 1.3 Sample data for example



Age	Occupation	Income	Policy Type	Policy Term	Premium Mode	Premium Amount
35	Employee	15,000/-	Life Security	25	Quarterly	2500/-
25	Employee	10,000/-	Life Security	25	Half Yearly	3000/-
55	Employee	65,000/-	Tax Benefit	20	Monthly	6000/-
45	Employee	45,000//	Tax Benefit	20	Monthly	5000/-
40	Business	70,000/-	Investment	25	Yearly	75,000/-
35	Business	90,000/-	Investment	25	Yearly	1,00,000/-

This example assumes that the problem is to classify customers in terms of different policy attributes such as policy term, premium amount, premium mode and policy type. The policy type classification can simply be done using income as main criteria shown below

10,000 ≤ Income ≤ 40,000	Life Security
45,000 ≤ Income ≤ 70,000	Tax Benefit
Income ≥ 70,000	Investment

The policy term require complicated set of divisions using both Age and Occupation. Similarly premium mode require complicated set of divisions using both Income and Occupation while premium amount require much more complicated set of divisions using Age, Income and Occupation

> DEFINITION 1.1. Given a database D = {t1,t2,…,tn } of tuples (items, records) and a set of classes C = {C1,…,Cm}, the classification problem is to define a mapping f : D → C where each ti is assigned to one class. A class Cj, contains precisely those tuples mapped to it that is, Cj = { ti | f(ti) = Cj, 1≤ i ≤ n, and ti Є D}

Algorithm 1.3 K Nearest Neighbors

When classification is to be made for new item using K Nearest Neighbors algorithm, its distance to each item in the training set must be determined. The new item is then placed in the class that contains the most items from the (K) closest set.

```
 Input:
        T   //Training data
        K   //Number of neighbors
        t   //Input tuple to classify
Output:
        c  //class to which t is assigned
Algorithm:
        N = ø
//Find the set of neighbors, N,  for t
        For each d Є T do
        If |N|≤ K, then
          N = N Ụ {d};
        else
          if     u Є N ∃such that sim(t,u) ≤ sim(t,d), then
               begin
                  N = N – {u};
                  N = N Ụ {d};
               end
                  //Find class for classification
C=class to which the most u Є N are classified;
```

For example, for life security policy there can be two groups as first is for customer with age 25 - 35, Income 10000/- to 15000/- and Occupation Employee with policy term of 20 years, premium mode quarterly and 16% premium amount of their income. Similarly second one is for customer with age 20 - 25, Income 5000/- to 10000/- and Occupation Employee with policy term of 25 years, premium mode half yearly and 30% premium amount. Suppose a customer with age 34, occupation Employee and Income 14000 purchasing Life security policy, will be suitable for first class i.e. customer with age 25 - 35, Income 10000/- to 15000/- and Occupation Employee can be suggested policy term of 25 years, premium mode quarterly and premium amount of 2200/-

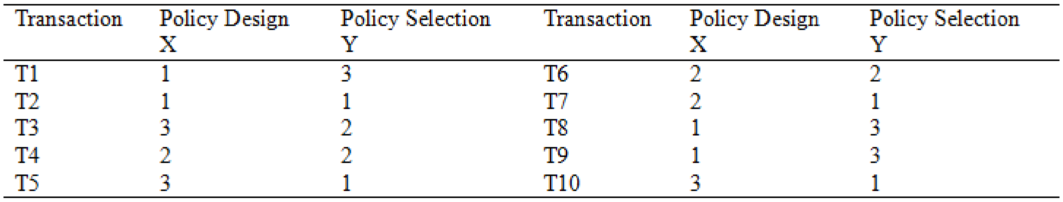
1.4 Correlation between Policy designing and policy selection

While studying policy designing factor and policy selection factor as a two variables simultaneously for a fixed population, insurance company can learn much by displaying bivariate data in a graphical from that maintains the pairing. Such pair wise display of variables is called a scatter plot. When there is an increasing trend in the scatter plot, we say that the variables have a positive association. Conversely, when there is a decreasing trend in the scatter plot, we say that the variables have a negative association. If the trend takes shape along a straight line, then we say that there is a linear association between the two variables.

Example 1.4

Insurance companies can consider the population consisting previous policy holders, and can investigate whether customers tend to purchase policy for the cause for which it is designed. To address this question, company needs to look at pairs of policy designing factor and policy selection factors.

Table 1.4 Sample data for example



Transaction	Policy Design X	Policy Selection Y	Transaction	Policy Design X	Policy Selection Y
T1	1	3	T6	2	2
T2	1	1	T7	2	1
T3	3	2	T8	1	3
T4	2	2	T9	1	3
T5	3	1	T10	3	1

n=10

To do so we can assign numbers to different policy designing and selection factors such as 1 for life security, 2 for investment and 3 for tax benefit etc. Then while analyzing previous policies we can put a respective numbers both for policy designing and selection factors and from that bivariate data, we can find increasing or decreasing trends between two factors.

Consider the population consisting of purchasing transactions, and we want to investigate whether people tend to purchase policy for the reason for which it is designed. Going a sample size of n and bivariate data set on these individuals or objects, the strength and linear relationship between the two variables X and Y is measured by the sample correlation coefficient r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

The mathematical formula for computing r is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Where n is a sample size

The value of r is such that $-1 < r < +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.

2. CONCLUSION

In the insurance industry, data mining can help firms gain business advantage mainly to support decision making. The insurance company needs to know the essentials of decision making and data mining techniques to compete in the market of life insurance. An understanding of probability and statistical distributions is necessary to absorb and evaluate acquiring new customers, retaining existing customers, performing sophisticated classification and correlation between policy designing and policy selection. Clustering technique can be used to acquire new customers in which first cluster specifies the group of customers holding life security policy while second group holds customer for Tax benefit policy and third group is for those customers holding policy for investment. When an agent approaches a particular customer, the agent will enter the demographic data of that customer in terms of age, occupation, income and education. Then each individual factor is compared with means of each cluster and the difference will be calculated. After comparing the each difference for each group, the closest cluster will be finalized which has the least difference. Association rule can be used to retain existing customers in which by reviewing previous data and by finding the required combinations according to confidence and support, agent can sell new policies to the existing customers to retain them. Similarly the company can also design such combo plans for their customer with additional benefits. Classification can be used to targeting customers or designing new products. Normally classes can be created according to policy term, premium mode and premium amount based on age, income and occupation. Policy term can be decided according to age and occupation while premium mode and premium amount can be can be according to income and occupation. So particular class for particular customer can be created where policy term, premium mode and premium amount can be mentioned in it in terms of percentage. Same way correlation can be used to identify the relation between policy designing and selection factors. To do so we can assign numbers to different policy designing and selection factors such as 1 for life security, 2 for investment and 3 for tax benefit etc. Then while analyzing previous policies we can put a respective numbers both for policy designing and selection factors and from that bivariate data, we can find increasing or decreasing trends between two factors. It is no wonder that the general insurance actuary must be a practicing statistician to gain a greater understanding of their business to help reduce fraud, improve underwriting and enhance risk management.

REFERENCES

[1] Alex Berson and Stephen J. Smith, “Data Warehousing, Data Mining, And OLAP”,MC Graow–Hill, 1997.

[2] Bigus and Joseph P, “Data Mining With Neural Networks”, MC Graw–Hill, New York 1996.

[3] Christopher J. Matheus, Gregory Piatetshy–Shapiro and Dwight Mcneill”, Selecting and Reporting what is Interesting The Kefir Application to Health Care Data”, Advances in Knowledge Discovery and Data Mining, AAA1 Press/The MIT Press, 1996.

[4] Dasrathy B. V., Ed, “Nearest Neighbor Norms: NN Pattern Classification Techniques”,IEEE, Computer Society Press, Calif. 1990.

[5] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, “Efficient Mining of Association Rules in Distributed Databases”, IEEE, 1996.

[6] Graig Silverstein, Sergey Brin and Rajeev Montwani, “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules”, Data Mining and Knowledge Discovery, Vol. 2, No. 1, Jan 1998, Kluwer Academic Publishers.

[7] Hongjun LU, Ling Feng and Jiawei Han, “Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules”, ACM Transactions on Information Systems, Vol. 18, October 2000.

[8] Huan Liu, Farhad Hussain, Chew Lim Tan and Manoranjan Dash, “Discretization: An Enabling Technique", Data Mining and Knowledge Discovery”, vol. 6 No. 4, October 2002.

[9] J. Date, "An Introduction to Database Systems", Addition Wesley Longman, Seven Edition, 2000.

[10] Jiawei Han, Laks V. S. Lakshmanan and Raymond T.NG, “Constraint-Based Multidimensional Data Mining”, IEEE, August 1999.

[11] Jorg-Uwe Kietz, Regina Zucker and Anca Vaduva, “Mining Mart: Combining Case- Based-Reasoning and multi-Strategy Learning Into a Frame For Reusing KDD-Applications”, Proc 5th Workshop on Multi-Strategy Learning (MSL 2000) Portugal, June 2000, Kluwer Academic Publishers.

[12] Ken Orr, “Data Warehouse Technology”, Copyright. The Ken Or Institute, 1997.

[13] Krzysztof J. Cios, Witold Pedryez and Roman W. Surniarski, “ Data Mining Methods for Knowledge Discovery“, Kluwer Academic Publishers 1998 Second Printing 2000.

[14] Mariano Fernendez Lopez, Asuncion Gomez-Perez, Juan Pazos Sierra, Polytechnic and Alejandro Pazos Sierra, “Building a Chemical Ontology Using Methontology and the Ontology Design Environment”, IEEE Intelligent System. Jan / Feb 1999.

[15] Martin Staudt, Anca Vaduva and Thomas c, “Metadata Management and Data Warehouse", Technical Report, Information System Research, Swiss Life, University of Zurich, Department of Computer Science, July 1999. vaduva@ifi.unizh.ch

[16] Ming-Syan chen, Jiawei Han and Philip S. Yu, “Data Mining: An Overview From a Database Perspective”, IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.

[17] Natalya Friedman Noy and Carole D. Hafner, “The State of The Art in Ontology Design”, AI Magazine Vol. 18, No. 3, Fall 1997.

[18] Rakesh A. grawal, “Parallel Mining of Associations Rule”, IEEE, Dec 1996.

[19] Ramakrishnan Srikant and Rakesh A. Grawal, “Mining Quantitative Association Rules in Large Relational Tables”, Proc Sigmod ‘96, 6/96 Montreal Canada, 1996 ACM.

[20] Ramakrishnan Srikant and Rakesh A. Grawal, “Mining Generalized Association Rules", Proceedings of The ‘21st VLDB Conference”, Zurich, Switzerland, 1995.

[21] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Hon and Alex Pany, “Exploratory Mining and Pruning Optimizations of Constrained Associations Rules”, ACM 1998 page 13.

[22] Mr. A. B. Devale and Dr. R. V. Kulkarni “A REVIEW OF DATA MINING TECHNIQUES IN INSURANCE SECTOR” Golden Research Thoughts Vol - I , ISSUE - VII [January 2012]

Authors

1) Amol B. Devale.

Working as Assistant Professor at Arts, Commerce and Science College, Palus, (Maharashtra) India. He has 10 years teaching experience in computer science and completed M..Phil degree at SIBER, Kolhapur (India). His research areas of interest are Data mining, expert system.

2) Dr. R. V. Kulkarni.

has professor at Chh. Shahu Institute of Business and Research, Kolhapur (India) and 30 years teaching experience. He has published more than 30 international and 50 national publications. Under his guidance 4 researcher completed Ph. D and 5 researcher are currently do Ph.D. . He is IT Consultant for sugar industry and Co-operative banks.