ISSN 2250-3153

Unsupervised Learning in Large Datasets for Intelligent Decision Making

S.Balaji[1], Dr.S.K.Srivatsa[2]

[1]Research Scholoar,Vels University Email: srisaibalaji@rediffmail.com
[2]Senior Professor, St.Joseph Eng.College Chennai-600100

Abstract- Unsupervised learning is often derived from motivations that appear to be independent of supervised learning. While supervised models aim at prediction, unsupervised models are mainly used for grouping records or fields and for the detection of events or attributes that occur together. Predictive analytics that predict the present are based on existing data, preferably as much as possible. Predictive techniques for customer segmentation can be used for intelligent decision towards customer preferences. In this paper unsupervised clustering based analysis towards very large datasets for analysis towards health insurance dataset for their preferences towards health insurance products for intelligent decision making.

Index Terms- Unsupervised learning,insurance,decision making,clustering

I. INTRODUCTION

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The most distinct characterstic of data mining is that it deals with large datasets (megabytes or even giga or terabytes).This requires the algorithms used in data mining to be scalable. Clustering is a popular approach used to implement unsupervised classification and data summation as well as segmentation of large heterogeneous data sets in to smaller homogeneous subsets that are easily managed,separately modeled and analysed.

II. BACKGROUND WORK

Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. For the insurance industry, there may be no such thing as an unacceptable consumer

Profile[4].Marisa s.viveros addressed unknown behaviour patterns from data collected in health information system.[5] Mittal & Kamakura (2001) find the link between customer satisfaction and retention to be moderated by customer characteristics.kanwal garg(2008) find decision tree method for identifying customer behaviour of investment in life insurance sector. Patrick A Rivers(2010) examined some of the benefits and challenges of using data mining processes within the health-care arena

III. METHODOLOGY

In unsupervised or undirected models there is no output field, just inputs. The pattern recognition is undirected; it is not guided by a specific target attribute.The goal of our model is to uncover data patterns in the set of input fields. Cluster model the groups are not known in advance. Instead the cluster algorithm to analyze the input data patterns and identify the natural groupings of records or cases. When new cases are scored by the generated cluster model they are assigned to one of the revealed clusters.

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. This algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Cluster based customer segmentation can target the profitable or valuable customers from other non-profitable customers.

IV. EXPERIMENTS

A portfolio database in an Health insurance company contains a set of insurance policies purchased by customers.our scope of analysis is limited to Health insurance policies.

Health insurance includes all risks related to the lives of human beings.A health insurance company's funds are collected by way of premiums.Every premium represents a risk that is covered by that premium.With the focus on dealing with large datasets data mining provides an environment in which to perform such extended analyses of health insurance portfolio databases.

The source dataset was extracted from IRDA Health insurance database.An intial trial set of some 15000 records covering a time period of one year(1st January 2010 to 31st January 2011).Each record contains 34 fields of data.

The first task was to remove from the database those variables which were irrelevant to the task at hand.This process was complicated by the fact that some obviously irrelevant attributes (eg:office at health insurance policy was taken).On the other hand,leaving irrelevant attributes in the data set can lead to aberrant results.

The second task labeled cleansing involved.It is the process of performing various transformations on data.(eg.Birthdates transformed into ages).

4.1 K-Means technique

Simple k-means algorithm is applied on Health Insurance data set for cluster formation and subsequent analysis to predict customer preferences towards health insurance.we considered 14 relevant attributes for analysis.

The set of 14 attributes after preprocessing taken for analysis is given in figure 4.1

Fig 4.1: Attributes of the dataset for the experiment and analysis

Slno	Attribute
1.	Txt_Gender
2.	age
3.	Txt_Type_of_Policy
4.	Boo_Pre-existing_Diseases_Covered
5.	Boo_Waiver_of_1st_Year_Exclusion
6.	Boo_Maternity_Cover
7.	Boo_Baby_cover_as_part_of_Maternity
8.	Boo_Floater_applicable
9.	Num_Sum_Insured
10.	Txt_Relationship_of_Insured
11.	Txt_Occupation
12.	Num_Individual_Premium
13.	Num_Floater_Amount
14.	Boo_Baby_cover_from_date_of_birth
15.	Txt_Gender

The weak tool has been considered for the purpose of analysis and test results. The WEKA ("Waikato Environment for Knowledge Analysis") tool is used for Data mining. Data mining finds valuable information hidden in large volumes of data. Weka is a collection of machine learning algorithms for data mining tasks, written in Java and it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The key features of Weka are it is open source and platform independent.

kMeans Model and evaluation on training set is given in figure 4.2

```
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 7278.247045117349
Missing values globally replaced with mean/mode

Cluster centroids:
Cluster#
Attribute Full Data      0          1          2          3          4
          (15000)      (3821)     (3315)     (3645)     (2010)     (2209)
=============================================================================
Txt_Gender     TRUE       FALSE      TRUE       TRUE       TRUE       TRUE

age           37.6765     42.7304    23.0609    40.7128    56.8303    28.4296

Txt_Type_of_Policy 3.8868   3.8307     3.9704     3.8132     3.9303     3.9402

Boo_Pre-existing_Diseases_Covered FALSE  FALSE    FALSE      FALSE   FALSE    FALSE

Boo_Waiver_of_1st_Year_Exclusion FALSE FALSE  FALSE FALSE      FALSE    FALSE

Boo_Maternity_Cover            TRUE    TRUE    TRUE     TRUE    TRUE       TRUE

Boo_Baby_cover_as_part_of_Maternity  TRUE   TRUE   TRUE  TRUE    TRUE      TRUE

Boo_Floater_applicable    TRUE     TRUE     TRUE       TRUE       TRUE    TRUE

Num_Sum_Insured 241137 201114.8914 356820.5128 203949.2455 333450.2488 114126.3015

Txt_Relationship_of_Insured 2.9869  3.7493   2.2449     2.3912     2.8055     1.5192

Txt_Occupation 6.2019    8.1874      6.524      5.2154     7.7442     2.5084

Num_Individual_Premium 3.2849     10.4316         0      2.5827   0          0

Num_Floater_Amount 241091.6667 201069.0919 356820.5128 203928.6694 33276.1194
114090.086
Boo_Baby_cover_from_date_of_birth  TRUE  FALSE   TRUE  FALSE  TRUE       TRUE

Time taken to build model (full training data) : 4.94 seconds
=== Model and evaluation on training set ===

Clustered Instances

0       3821 ( 25%)
1       3315 ( 22%)
2       3645 ( 24%)
3       2010 ( 13%)
4       2209 ( 15%)
```

Figure 4.2 Results of simple –k means clustering on dataset

The first column gives you the overall population centroid. The second and third columns gives the centroids for cluster 0 to 4 respectively. Each row gives the centroid coordinate for the specific dimension.

In the above results cluster 0 represents House wives -female gender falls with average age of 42 hold group policy with no pr-existing diseases.

In Cluster 1 there are male genders with average age of 23 will prefer for Individual policy.

In cluster 2 there are male genders with average age of 42 is of production workers prefers group policy.

Cluster 3 represents the male with an average age of 40 with high risk profile of occupation like police/defence/paramilitary

Prefer group floater policy.

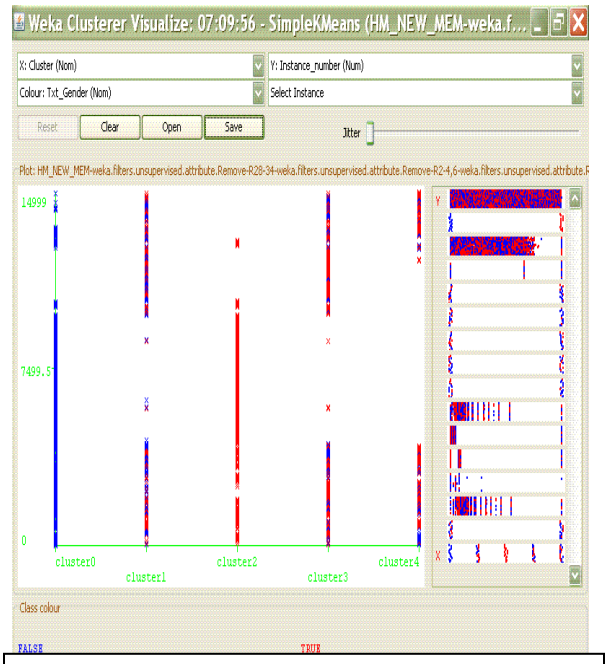
Cluster4 represents the male with average age of 28 are of business traders and professional prefer self policy rather than group policy.

Clustered Instances

0 3821 (25%)
1 3315 (22%)
2 3645 (24%)
3 2010 (13%)
4 2209 (15%)

The clustering model shows the centroid of each cluster and statistics on the number and percentage of instances assigned to different clusters. Cluster centroids are the mean vectors for each cluster; so, each dimension value and the centroid represents the mean value for that dimension in the cluster.

Thus, centroids can be used to characterize the clusters. Numbers are the average value of everyone in the cluster. Each cluster shows us a type of behavior in customers, from which conclusions are drawn.



4.3 output weka clustering visualization

The visualization of above set of data is projected in figure 4.3.

For the chosen data set the cluster number as the x-axis, the instance number (assigned by WEKA) as the y-axis, and the "gender" attribute as the color dimension. This will result in a visualization of the distribution of males and females in each cluster.The observered results as clusters 2, 3 and 4 are dominated by males, while clusters 0 and 1 are dominated by females. In this case, by changing the color dimension to other attributes, their distribution within each of the clusters can be observed.

5.2 J4.8 technique:-

Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. WEKA has implementations of numerous classification and prediction algorithms. The basic ideas behind using all of these are similar. In this example we will use the modified version of the bank data to classify new instances using the C4.5 algorithm (note that the C4.5 is implemented in WEKA by the classifier class: weka.classifiers.trees.J48).The Health insurance datset of 15,000 records was taken for analysis with 14 relevant attributes . The dataset is fed into weka's classification analyzer with J4.8 Rule set. The result of the run information is given in figure 4.4.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Txt_Occupation <= 1: TRUE (5410.0/860.0)
Txt_Occupation > 1
|   Txt_Occupation <= 8: FALSE (4829.0)
|   Txt_Occupation > 8
|   |   Txt_Relationship_of_Insured <= 5: TRUE (2464.0/12.0)
|   |   Txt_Relationship_of_Insured > 5
|   |   |   Txt_Relationship_of_Insured <= 6: FALSE (1259.0)
|   |   |   Txt_Relationship_of_Insured > 6
|   |   |   |   Boo_Waiver_of_1st_Year_Exclusion = TRUE: FALSE (2.0)
|   |   |   |   Boo_Waiver_of_1st_Year_Exclusion = FALSE
|   |   |   |   |   Num_Sum_Insured <= 350000: FALSE (13.0/3.0)
|   |   |   |   |   Num_Sum_Insured > 350000: TRUE (23.0/10.0)

Number of Leaves  :     7

Size of the tree :     13


Time taken to build model: 2.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       14110               94.0667 %
Incorrectly Classified Instances       890                5.9333 %
Kappa statistic                          0.8798
Mean absolute error                      0.0996
Root mean squared error                  0.2234
Relative absolute error                 20.0205 %
Root relative squared error             44.7779 %
Coverage of cases (0.95 level)          99.88   %
Mean rel. region size (0.95 level)      68.1333 %
Total Number of Instances           15000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
Class
                 0.874     0.001     0.999     0.874     0.932      0.962
FALSE
                 0.999     0.126     0.901     0.999     0.947      0.962
TRUE
Weighted Avg.    0.941     0.068     0.946     0.941     0.94       0.962

=== Confusion Matrix ===

    a    b   <-- classified as
 6099  883 |    a = FALSE
    7 8011 |    b = TRUE
```

Figure 4.4.Output from WEKA's classification model

An attribute is chosen to compare clusters and classes defined by the values of this attribute. The evaluation is done by

displaying the confusion matrix.The gender attribute is used for classifying the classes.

The important numbers to focus on here are the numbers next to the "Correctly Classified Instances" (94.0667 percent) and the "Incorrectly Classified Instances" (5.9333 percent). Other important numbers are in the "ROC Area" column, in the first row (the 0.962).

Comparing the "Correctly Classified Instances" from this test set (94.0667 percent) with the "Correctly Classified Instances" from the training set (5.9333 percent), it is observed that the accuracy of the model is good enough to predict the gender based classification of customers towards health insurance policy .

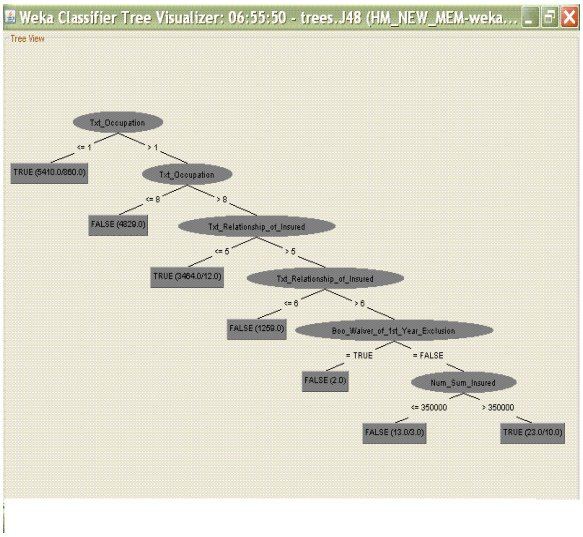


Fig 5.5 Decision Tree construction using J4.8 algorithm

From the above decision tree figure 5.5 generated from J4.8 algorithm, it is observed that the occupation attribute of the policy holder has the significant impact inaddition to the gender and age classification towards group policy or self policy preference of health insurance.

V. Conclusion

Analysis of policyholders behavior enables companies to improve support of their policy holder oriented processes, which aims to improve the overall performance of the health insurance company. Unsupervised data mining methodology has a tremendous contribution for researchers to extract the hidden knowledge and information [12].customer segmentation generated by clustering based model will add value by enabling target appropriate products to different consumers. The research described in this paper also identified gender based preferences towards health insurance policies in addition other attributes.

References

[1] E.W.T. Ngai , Li Xiu and D.C.K. Chau, 2009, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, Vol 36,Issue 2, Part 2 , pp 2592-2602.

[2] Zhexue Huang,A fast culstrering algorithm to cluster very large categorical data sets in Data mining,Research paper for Advanced Computational systems,under Australian governments cooperative research centers program

[3] zhiyuan yao,Annika H.Holmbom,Tomas Eklund,Combining Unsupervised and supervised data mining techniques for conducting customer portfolio analysis,Springer-Verlag Berlin Heidelberg 2010.,ICDM 2010 pp292-307.

[4] Hossack, I.B. et al., eds. (1999), Introductory Statistics with Applications in General Insurance, Cambridge University Press.

[5] Marisa s.viveros,John P.Nearhos,Michael J.Rothman,Applying data mining techniques to a Health Insurance information system,proceedings of the 22nd VLDB conference,1996.

[6] Zhiyuan Yao,Annika H.Holmbom,Tomas Eklund and Barbro Back,Combining Unsupervised and Supervised Data Mining Techniques for Conducting Customer Portfolio Analysis,ICDM 2010.LNAI 6171,pp.292-307,2010

[7] Khurana Sunayna. (2008)."Customer references in Life Insurance Industry in India", The ICFAIUniversity Journal of Services vol. 6(3), 61-68.

[8] Chien-Hsing Wu,, Shu-Chen Kao, Yann-Yean Su, Chuan-Chun Wu, Targeting customers via discovery knowledge for the insurance industry, 0957-4174/$ - see front matter q 2005 Elsevier Ltd.

[9] Hokey Min, Developing the Profiles of Supermarket Customers through Data Mining[J], The Service Indusries Journal, Vol.26, No.7, October 2006, pp.747-763

[10] Seyed Mohammad Seyed Hosseini ,Anahita Maleki, Mohammad Reza Gholamian, 2010, Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty, Expert Systems with Applications, Vol 37, Issue 7, pp. 5259-5264.

[11] Shim, Beom-Soo and Suh, Yong-Moo, 2010, CRM Strategies for A Small-Sized Online Shopping Mall Based on Association Rules and Sequential Patterns, *PACIS 2010 Proceedings.*

[12] Ismail, R., Othman, Z. and Bakar, A.A,2010, Associative prediction model and clustering for product forecast data, Intelligent Systems Design and Applications(ISDA), 10th International Conference

Authors

First Author – S.Balaji, Research Scholoar,Vels University
Email: srisaibalaji@rediffmail.com
Second Author – Dr.S.K.Srivatsa, Senior Professor, St.Joseph Eng.College Chennai-600100