Cluster Analysis in Data Mining using K-Means Method

[1]Narander Kumar
Department of Computer Science
B. B. Ambedkar University
Lucknow (U.P.), 226025,INDIA

[2]Vishal Verma
Department of Computer Science,
B. B. Ambedkar University
Lucknow (U.P.), 226025,INDIA

[3]Vipin Saxena
Department of Computer Science,
B. B. Ambedkar University
Lucknow (U.P.), 226025,INDIA

ABSTRACT

To find the unknown and hidden pattern from large amount of data of insurance organizations. There are strong customer base required with the help of large database. Cluster Analysis is an excellent statistical tool for a large and multivariate database. The clusters analysis with K-Means method may be used to develop the model which is useful to find the relationship in a database. In this paper, consider the data of LIC customer, the seeds are the first three customers then compute the distance from cluster using the attributes of customers with the help of Clustering with K-Means method. Comparing the mean distance of cluster with the seeds. Finally, we find the nigh distances from the cluster as the cluster (C1) have three customers named S1, S2, S10 which are satisfy with all the benefits, terms and conditions of cluster (C1). If requirements of any customer same as the S1, S2, S10 then we allocated the cluster (C1). It will increase the revenue as well as profit of the organization with customer satisfaction.

Keywords K-means methods, Seeds, Clustering analysis, Cluster distance, LIPS.

1. INTRODUCTION

Cluster analyses have a wide use due to increase the amount of data .k-means method is a technique for analysis. The aim of cluster analysis is to find the optimal division of m entities into n cluster of K-means cluster analysis is eg. Insurance Policy System. The data is taken from the Life Insurance Corporation of India. Through the cluster analysis method is that grouped the data objects into cluster based on method to represents the same type of data objects as well as according to requirements of application while data-based methods represent actual structure of data by representing the similar data objects together. Clustering methods as an optimization problem try to find the approximate or local optimum solution. The advantage of the cluster likewise in a medical science, when a doctor take three or two people in a group they show their symptoms and they find out their disease. When they know the symptoms of various diseases by such type of analysis of people's symptoms and when another person come to doctor show the symptoms, the doctor tell him straight forward the name of disease. In this case they need not to check the person because of analysis.

In this paper, the aim to find out any interesting group of the persons who wants to take policy based on some conditions according to the policy or who satisfy their needs. Therefore through cluster analysis using K-Means method, find out the group of persons who belongs to same criteria which is affordable the premium of policy and satisfy with the other benefits. Consider the data about the Life Insurance Corporation in which we consider first three people as the three seeds (named as S1, S2, and S3 customers) for finding out the cluster using K-Means method. Compute the distance using the attributes and use sum of differences. Based on these distance each person is allocated to nearest cluster. Now compare the cluster means with original seeds use the new cluster means to recompute the distance. When the cluster shows that the cluster have not change specify them as the final cluster.

2. RELATED WORK

Nowadays clustering techniques have wide use to group the data in to same type of objects. K-means is a technique for clustering analysis using above said techniques. Give an analysis of crop yield record by Weka Interface [1], they also included analysis of rice data after demonstration via Weka Interface .A concept based clustering approach find genes and proteins that have a similar functionality has been given in [2]. A voice activity detection algorithm based on spectral clustering methods which divides the input signal into two clusters.ie speech presence and absence frames. Using these data clusters, they apply Laplacion and Gaussian models to compute the ratio needed for voice activity detection has been proposed in [3]. A solution through algorithm for the problem of finding a partition maximizing the modularity of a given graph can be reduced to a minimum weighted cut problem on a complete graph with the same vertices has been given in [4] formulation of co-clustering as a constrained multi-linear decomposition with sparse latent factors has been given in [5]. A comparison of neighborhood constrained clustering algorithm and advance spectral spatial clustering algorithms has been given in [6]. An innovative probabilistic clustering concept for aggregate modeling of wind farms has been proposed in [7].

The importance and significance of data mining techniques and discussed management tools for customer relationship management through finding the hidden information for a insurance company has been given in [8]. An algorithm [9] for building decision model for insurance with health and wealth of investment.Seenario discovery methods which use statistical data mining algorithm and they identify that assumptions and system conditions that are affected the cost in [10]. Cluster analysis is present in [11] to analyse the things for designing the catalogue and cluster's consumption preference. An investigation is given [12] the functionalities which are according to need of consumers and wants to take an insurance product by extracting pattern of knowledge. Comparisons of three systems such as cover story explore and KDW has been given [13]. A survey for review the methods for discovery of domain –specific query form is given [14].Importance and significance of data mining technique

and discussed about the tools which help to make a good

C1	22	45	13	14	From C1	From C2	From C3	Allocation to the nearest cluster
C2	22	15	25	13				
C3	34	9	60	19				
S1	22	45	13	14	0	39	57	C1
S2	22	15	25	13	43	0	59	C2
S3	34	9	60	19	100	50	0	C3
S4	37	7	30	24	80	39	40	C2
S5	35	15	42	33	91	50	49	C3
S6	38	18	28	36	70	35	62	C2
S7	32	7	36	75	132	75	84	C2
S8	45	15	29	66	121	80	95	C2
S9	44	15	14	90	132	110	133	C2
S10	24	30	20	30	46	45	88	C2

customer relationship management for insurance company in

Customer	Age	Policy-premium 1	Policy-premium2	Policy-premium3
S1	22	45	13	14
S2	22	15	25	13
S3	34	9	60	19
S4	37	7	30	24
S5	35	15	42	33
S6	38	18	28	36
S7	32	7	36	75
S8	45	15	29	66
S9	44	15	14	90
S10	24	30	20	30

[15]. Theoretical framework for studying e-business value is given [16].Survey of ambient intelligent with its applications has been discussed in [17].

There are a wide use and importance of clustering techniques using K-Means methods, we use these techniques on the Life Insurance Policy System. In this paper, there are consider the data of Life Insurance Policy System in which the attributes are the age and the three policy premium. First three people as the three seeds (named as S1, S2, and S3 customers). We analyze the distance from cluster and compare the cluster mean of cluster with the original seed by clustering analysis using K-Means method. According to distance, each customer is allocated to nearest cluster. Compare the cluster means with original seeds use the new cluster means to recompute the distance. When the cluster shows that the cluster have not change specifically them as the final cluster.

3. ANALYSIS OF CLUSTER

The K-Means Method

Consider the Data about LIC customer in given table. The only attributes are the age and the three policy premium.

Table 1: Data for Policy Premium rupee in hundred

Distance from cluster

Let the tree seeds be the first tree customer as shown in table 1

Customer	Age	Policy-premium1	Policy-premium2	Policy-premium3
S1	22	45	13	14
S2	22	15	25	13
S3	34	9	60	19

The first iteration lead 1customer in first cluster & 2 customers in second cluster & 3 customers in third cluster.

Comparing the cluster mean of cluster with the original seed

	Age	Policy-premium1	Policy-premium2	Policy-premium3
C1	22	45	13	14
C2	34.57	15.28	23.428	48.57
C3	15.28	12	51	26
Seed1	22	45	13	14
Seed2	22	15	25	13
Seed3	34	9	60	19

Now step 3

Use the new cluster means to evaluated the distance of each object to each of the mean again allocating each object to the nearest cluster

C1	22	45	13	14	From C1	From C2	From C3	Allocation to the nearest cluster
C2	34.57	15.28	23.42	48.57	0			
C3	34.5	12	51	26				
S1	22	45	13	14	0	75.28	95.5	C1
S2	22	15	25	13	43	49.44	54.5	C1
S3	34	9	60	19	70	59.3	19.5	C3
S4	37	7	30	24	69	41.86	30.5	C3
S5	35	15	42	33	91	34.86	19.5	C3
S6	38	18	28	36	80	23.3	42.5	C2
S7	32	7	36	75	122	49.86	71.5	C2
S8	45	15	29	66	121	33.72	75.5	C2
S9	44	15	14	90	139	60.56	113.5	C2
S10	24	30	20	30	40	47.28	63.5	C1

	Age	**Policy-premium1**	**Policy-premium2**	**Policy-premium3**
C1	22.66	30	19.33	19
C2	39.7	13.75	26.75	66.75
C3	35.33	10.33	44	25.33
Seed1	22	45	13	14
Seed2	34.57	15.28	23.42	48.57
Seed3	34.5	12	51	26

C1	**22.66**	**30**	**19.33**	**19**	**From C1**	**From C2**	**From C3**	**Allocation to the nearest cluster**
C2	39.7	13.75	26.75	66.75				
C3	35.33	10.33	44	25.33				
S1	22	45	13	14	26.90	115.45	90.33	C1
S2	22	15	25	13	27.33	74.45	49.33	C1
S3	34	9	60	19	73.01	91.45	24.99	C3
S4	37	7	30	24	53.01	55.45	20.33	C3
S5	35	15	42	33	64.01	54.95	16.33	C3
S6	38	18	28	36	53.01	37.95	37.01	C2
S7	32	7	36	75	75.01	31.95	64.33	C2
S8	45	15	29	66	94.01	9.55	70.01	C2
S9	44	15	14	90	112.67	41.55	108.01	C2
S10	24	30	20	30	13.67	71.45	65.67	C1

C_1---- S_1, S_2, S_{10}

C_2 --- S_6, S_7, S_8, S_9

C_3—S_3, S_4, S_5

From the K-Means Method we find the cluster (C1) have three customers named S1, S2, S10 which are the satisfy with all the benefits, terms and conditions of cluster (C1). If requirements of any customer same as the S1, S2, S10 then we allocated the cluster C1. Cluster C2, C3 allocated as the cluster C1.

4. CONCLUSIONS:

The increasing demand of information, which will help to policy makers for making strong customer relationship and good image in mind of customers. Clustering analysis with K-means method may set a path towards make a good relationship between customers and insurance policy organization and provide the customers satisfaction also. If customers are increased in any policy system then organization or company will also grow. In this paper, find out the mean distance of clusters on the basis customer attributes with the help of analysis of clustering. Through which, such customer have same requirements then allocate the nearest cluster. For future direction, other techniques of statistical analysis tool may be used to find out the distance between same clusters with help customer attributes.

5. REFERENCES

[1] Ritu Sharma, M. Afshar Alam, Anita Rani," K-Means Clustering in Spatial Data Mining using Weka Interface", International Conference on Advances in Communication and Computing Technologies (ICACACT) Proceedings published by International Journal of Computer Applications® (IJCA) 2012.

[2] Hurtado, C., Mendelzon, A. and Vaisman, A., Maintaining Data Cubes under Dimension Updates, Proc IEEE/ICDE '99 Ankit Gupta ,Arpit Gupta Amit Mishra, "RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS", International Journal of Advance Technology & Engineering Research (IJATER), ISSN NO: 2250-3536 ,Vol. 1, Issue 1, November 2011.

[3] Mousazadeh, S.,Cohen,I.,"Voice Activity Detection in Presence of Transient Noise Using Spectral Clustering", Audio, Speech, and Language Processing, IEEE Transactions on, Volume: 21 , Issue: 6,pp- 1261 – 1271, June 2013.

[4] Djidjev, Hristo N., Onus, Melih; "Scalable and Accurate Graph Clustering and Community Structure Detection", Parallel and Distributed Systems, IEEE Transactions on, Volume: 24 , Issue: 5,pp-1022 - 1029 , 26 March 2013 .

[5] Papalexakis, E.E., Sidiropoulos, N.D.; Bro, R.; "From K-Means to Higher-Way Co-Clustering: Multilinear Decomposition With Sparse Latent Factors", Signal Processing, IEEE Transactions on, Volume: 61 , Issue: 2 ,pp-493 – 506, Jan.15, 2013.

[6] Shanshan Li, Bing Zhang; An Li; "Xiuping Jia; Hyper spectral Imagery Clustering With Neighborhood Constraints", Geoscience and Remote Sensing Letters, IEEE Transactions on, Volume: 10 , Issue: 3 ,pp- 588 – 592, 23 November 2012.

[7] Ali, M., Ilie, I.-S.; Milanovic, J.V.; Chicco, G., "Wind Farm Model Aggregation Using Probabilistic Clustering Power Systems", IEEE Transactions on, Volume: 28 , Issue: 1 ,pp- 309 – 316, Feb. 2013.

[8] Jayanti Ranjan, Raghuvir Singh,Vishal Bhatnagar," Analytical customer relationship management in insurance industry using data mining: a case study of Indian insurance company" International Journal of Networking and Virtual Organizations, Volume 9 Issue 4, Pages 331-366, November 2011.

[9] Yu-Ju-Lin,Chin-Sheng, Huang,Che-Chern Lin, "Determination of insurance policy using neural networks and simplified models with factor analysis technique", International Journal WSEAS Transactions on Information Science and Applications, Volume 5 Issue 10, Pages 1405-1415, October 2008.

[10] Joseph R. Kasprzyk, Shanthi Nataraj, Patrick M. Reed, "Many objective robust decision making for complex environmental systems undergoing change" International Journal Environmental Modeling & Software, Volume 42, Pages 55-71, April 2013.

[11] Shu-hsien Liao, Yin-ju Chen, Yi-tsun Lin, "Mining customer knowledge to implement online shopping and home delivery for hypermarkets" International Journal Expert Systems with Applications, Volume 38 Issue 4, Pages 3982-3991, April 2011.

[12] Shu-Hsien Liao,Ya-Ning Chen, Yu-Yia Tseng, "Mining demand chain knowledge of life insurance market for new product development" International Journal Expert Systems with Applications, Volume 36 Issue 5, Pages 9422-9437, July 2009.

[13] C.J. Matheus,P.K. Chan,G. Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases" IEEE Transactions on Data Engineering, vol. 5 no. 6, pp. 903-913, December 1993.

[14] Mauricio C. Moraes, Carlos A. Heuser, Viviane P. Moreira," Pre-Query Discovery of Domain-specific Query Forms: A Survey" IEEE Trasactions on Data Engineering, ISSN: 1041-4347, , 22 May 2012.

[15] Vishal Bhatnagar, Jayanthi Ranjan, "Time to implement data mining in insurance firms for effective CRM and CRM analytics" International Journal of Networking and Virtual Organizations, Volume 9 Issue 1, Pages 1-24, June 2011.

[16] Kevin Zhu, Kenneth L. Kraemer, Jason Dedrick," Information Technology Payoff in E-Business Environments: An International Perspective on Value Creation of E-Business in the Financial Services Industry" International Journal of Management Information Systems, Volume 21 Issue 1, Pages 17-54, Number 1/Summer 2004.

[17] Fariba Sadri ," Ambient intelligence: A survey" International Journal ACM Computing Survey(CSUR), Volume 43 Issue 4, Article No. 36, October 2011.

BRIEF BIOGRAPHY

Dr. Narander Kumar received his Post Graduate Degree and Ph. D. in CS & IT, from the Department of Computer Science and Information Technology, Faculty of Engineering and Technology, M. J. P. Rohilkhand University, Bareilly, Uttar Pradesh, INDIA in 2002 and 2009, respectively. His current research interest includes Quality of Service (QoS), Software Engineering, Computer Networks, Resource Management Mechanism, in the networks for Multimedia Applications, Performance Evaluation. Presently he is working as Assistant Professor, in the Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, INDIA.

Vishal Verma is a research scholar in Department of Comuter Science, Babashaheb BhimRao Ambedkar University, Lucknow, India. Earlier he got his Master of Computer Application (MCA) from the above University and presently he is working on Data Mining Applications through UML.

Vipin Saxena is a Professor and Head, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He got his M.Phil. Degree in Computer Application in 1992 & Ph.D. Degree work on Scientific Computing from University of Roorkee (renamed as Indian Institute of Technology, Roorkee, India) in 1997. He has more than 16 years of teaching experience and 19 years of research experience in the field of Scientific Computing & Software Engineering. He has published more than ninety one International and National research papers and authored four books in the Computer Science field. Dr. Saxena is a life time member of Indian Science Congress.