International Journal of Advanced Computer Science and Information Technology (IJACSIT)
Vol. 6, No. 3, 2017, Page: 31-39, ISSN: 2296-1739

www.elvedit.com

A Data Mining Approach for Analysis of Customer Behavior in order to Improve Policies in Insurance Industry based on Combination of Particle Swarm Optimization and k-Means Algorithm

Authors

Hamid Moradi
Department of Computer Engineering/ Islamic Azad University, Qeshm Branch, Iran
hamidmoradi1392@yahoo.com
Qeshm, Iran

Javad Hosseinkhani
Department of Computer Engineering/ Islamic Azad University, Qeshm Branch, Iran
jhkhani@gmail.com
Qeshm, Iran

Abstract

One of the important branches of insurance products is individual insurance policies, which is directly cover insurer's life. Among individual insurance policies, health insurance policies because of diversity of coverages and contracts and high loss ratio are very important in insurance company. Due to high competition in the sale of policies between insurance companies and climb increasing of demand for them, knowing the customers of these products is considered important in the maintenance and survival of insurance organizations. This study is considering the practical application of data mining in an insurance company on health insurance policy customers to investigate whether in this way can help insurance companies to identify different customer groups and their characteristics in order to make suitable patterns for offering suitable services to customers. In this way the maximum value of the relationship with the customer is achieved. In this research, the customers of health insurance policies have been clustered by means of some features. The Clustering was done using proposed algorithm based on PSO and k-Means algorithms. Evaluation has shown that the proposed method has high accuracy in data clustering. The proposed model has clustered data in four clusters which each cluster differ from others in terms of usefulness to the organization. The

result has shown that the third cluster is the most profitable and fourth cluster is the most harmful. According to the proposals made to each cluster, organization can maximize benefits from the relationship with its policyholders. The Clustering was done using proposed algorithm based on k-Means algorithm. The proposed model has clustered data in four clusters which each cluster differ from others in terms of usefulness to the organization.

Key Words

Data Mining Approach, Analysis of Customer Behavior, Policies in Insurance Industry, Particle Swarm Optimization

I. INTRODUCTION

In healthcare organization which in turn helpful for all the parties associated with this field. Data Mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pat tern and useful information from huge dataset. Various studies highlighted that Data Mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision [1]. In general, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process [2, 3]. According to Javad et al., the knowledge discovery process are structured in various stages whereas the first stage is data selection where data is collected from various sources, the second stage is pre - processing of the selected data , the third stage is the transformation of the data into appropriate format for further processing, the fourth stage is Data Mining where suitable Data Mining technique is applied on the data for extracting valuable information and evaluation is the last stage.

Clustering is an unsupervised learning method unlike the classification method which is generally viewed as a supervised learning technique. In other words, clustering groups the data based only on the information that is available in the dataset without any labels [4]. Clustering techniques can be generally classified into partitional methods, Hierarchical approaches, Density-based algorithms, Probabilistic methods, Grid-based methods, Graph theory, Model-based approaches and so on [5]. Many partitional algorithms are recently introduced which are based on the technique of Evolutionary programming that includes Genetic Algorithms (GA), evolved from the Darwinian Theory.

Partitional clustering algorithms determine the clusters in such a way that the similarity within the clusters is maximum and the dissimilarity between the clusters is minimum. Though the K-Means algorithm is the most widely used algorithm under the partitional clustering because of its easy implementation factor, it has certain limitations. It does not give efficient results with differently shaped clusters [6] and moreover, it arbitrarily converges to local optima. But the clustering performance can be improved in terms of accuracy by incorporating constraints [7].

II. LITERATURE REVIEW AND RELATED WORKS

Generally all the healthcare organizations across the world stored the healthcare data in electronic format. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate. Due to continuous increasing the size of electronic healthcare data a type of complexity is exist in it. In other words, we can say that healthcare data becomes very complex. By using the traditional methods it becomes very difficult in order to extract the meaningful information from it. But due to advancement in field of statistics, mathematics and very other disciplines it is now possible to extract the meaningful patterns from it. Data mining is beneficial in such a situation where large collections of healthcare data are available

Data Mining mainly extracts the meaningful patterns which were previously not known. These patterns can be then integrated into the knowledge and with the help of this knowledge essential decisions can becomes possible. A number of benefits are provided by the data mining. Some of them are as follows: it plays a very important role in the detection of fraud and abuse, provides better medical treatments at reasonable price, detection of diseases at early stages, intelligent healthcare decision support systems etc. Data mining techniques are very useful in healthcare domain. They provide better medical services to the patients and helps to the healthcare organizations in various medical management decisions. Some of the services provided by the data mining techniques in healthcare are: number of days of stay in a hospital, ranking of hospitals, better effective treatments, fraud insurance claims by patients as well as by providers, readmission of patients, identifies better treatments methods for a particular group of patients, construction of effective drug recommendation systems, etc [14]. Due to all these reasons researchers are greatly influenced by the capabilities of data mining. In the healthcare field researchers widely used the data mining techniques. There are various techniques of data mining. Some of them are classification, clustering, regression, etc.

Every medical information related to patient as well as to healthcare organizations is useful. With the help of such a powerful tool known as data mining plays a very important role in healthcare industry. Recently researchers uses data mining tools in distributed medical environment in order to provide better medical services to a large proportion of population at a very low cost, better customer relationship management, better management of healthcare resources, etc. It provides meaningful information in the field of healthcare which may be then useful for management to take decisions such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction etc., [15]. Dealing with the issues and challenges of data mining in healthcare [16] in order to predict the various diseases effective analysis of data mining is used [17].

Some articles from the year 2005-2014 are taken and are used for survey for this category. In [8], Taher Niknam and Babak Amiriv have proposed a hybrid evolutionary algorithm called FAPSO-ACO-K (Fuzzy Adaptive Particle Swarm Optimization – Ant Colony Optimization – K-Means) to solve the nonlinear partitional clustering problem. The performance of this algorithm was evaluated through several benchmark datasets. The simulation results showed that the performance of this algorithm was better than the other traditional algorithms such as PSO (Particle Swarm Optimization), ACO, Simulated Annealing and so on.

In [9], M.Arshad designed a clustering algorithm based on KEA (Key phrase Extraction Algorithm) to solve the problem of document clustering in traditional clustering technique. The Kea Bisecting K-Means clustering algorithm was used to extract the test documents from a large amount of text documents in an easy and efficient way. The clustering algorithm was applied in order to generate the clustering document based on the extracted keys. The documents were grouped into several clusters like in the Bisecting K-Means algorithm. The results and the performance showed a consistently good quality of clusters which demonstrated in turn that the Bisecting K-Means is an excellent algorithm for clustering a large number of documents.

Three modified conventional moving K-Means clustering algorithms have been recommended by Nor Ashidi Mat Isa et al, [10] for the application of image segmentation. These three algorithms are fuzzy moving K-Means, adaptive moving K-Means and adaptive fuzzy moving K-Means algorithms. Standard images and hard evidence on microscopic digital image were used to analyze these algorithms. The segmentation result was compared with the conventional K-Means, fuzzy C-Means and moving K-Means algorithms. By qualitative and quantitative analysis, it was proved that this algorithm was less sensitive to noise and also the problems such as dead centers, center redundancy and trapped center at local minima were avoided. It was also illustrated that the above three modified algorithms were suitable to implement consumer electronics products based on their simplicity and capability.

- Clustering Examples in Healthcare

Chen et al., proposed hierarchical K-means regulating divisive or agglomerative approach for better analyzing large micro-array data. It was reported that divisive hierarchical K-means was superior to hierarchical and K-means clustering on cluster quality as well as on computational speed. Apart from this, it was also mentioned that divisive hierarchical K-means establishes a better clustering algorithm satisfying researcher's demand [18].

Chipman et al., proposed the hybrid hierarchical clustering approach for analyzing microarray data [19]. In this research, the proposed hybrid clustering approach combines bottom-up as well as top-down hierarchical clustering concepts in order to effectively and efficiently utilizes the strength of both concepts for analyzing micro-array data. The proposed approach was built on a mutual cluster. A mutual cluster is a group of points closer to each other than to any other points. The research demonstrates the proposed technique on simulated as well as on real micro-array data.

Tapia et al. analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm. In this research, the main focus was on regeneration of protein-protein functional interactions from genomic data. In this research, the proposed algorithm can predicate the functional associations accurately by considering genomic data [20].

Soliman et al., proposed a hybrid approach for better analyzing the cancer diseases on the basis of informative genes. The proposed approach used the K-means clustering with statistical analysis (ANOVA) for gene selection and SVM to classify the cancer diseases. On the basis of experiments that were performed on micro-array data, it has been found that the accuracy of K-means clustering with the combination of statistical analysis was better [21].

Belciug et al., concluded that among hierarchical, partitional, and density based clustering, the hierarchical clustering was provided effective utilization of hospital resources and provided improved patient care services in healthcare [22].

Schulam et al., proposed a Probabilistic Subtyping Model (PSM) which was mainly designed in order to discovered subtypes of complex, systematic diseases using longitudinal clinical markers collected in electronic health record (EHR) databases and patient registries. Proposed model was a model for clustering time series of clinical markers obtained from routine visits in order to identify homogeneous patient subgroups [23].

III. LIMITATION OF K-MEANS DATA CLUSTERING APPROACH

K-means clustering has some of the limitations which need to get overcome. Several people got multiple limitations while working on their research with K-means algorithm. Some of the common limitations are discussed below.

- Outliers

It has been observed by several researchers that, when the data contains outliers there will be a variation in the result that means no stable result from different executions on the same data. Outliers are such objects they present in dataset but do not result in the clusters formed. Outliers can also increase the sum of squared error within clusters. Hence it is very important to remove outliers from the dataset. Outliers can be removed by applying preprocessing techniques on original dataset [11].

- Number of clusters

Determining the number of clusters in advance is always been a challenging task for K-means clustering approach. It is beneficial to determine the correct number of clusters in the beginning. It has been observed that sometimes the number of clusters is assigned according to the number of classes present in the dataset. Still it is an issue that on what basis the number of clusters should be assigned [12].

- Empty clusters

If no points are allocated to a cluster during the assignment step, then the empty clusters occurs. It was an earlier problem with the traditional K-means clustering algorithm [11].

Non globular shapes and sizes

With the K-means clustering algorithm if the clusters are of different size, different densities and non-globular shapes, then the results are not optimal. There is always an issue with the convex shapes of clusters formed [13].

IV. EVALUATION OF PROPOSED METHOD

In order to evaluate the proposed method in this research, the standard Benchmark data of UCI machine learning data has been used. The selected data sets were measured using the proposed method and the clustering

method using PSO clustering and then the silo values for each method in the number of different clusters. The characteristics of the standard data set used in the UCI data warehouse are as follows:

TABLE I. FEATURES OF THE DATA SET USED

Number of records	Number of features	Dataset Name
210	7	Seeds
403	5	User Knowledge Modelling
150	4	Iris
214	10	Glass

The proposed algorithm has achieved more favorable results than the PSO clustering algorithm available in the research literature:

TABLE II. THE MEAN VALUES OF THE SILO CRITERION FOR EACH DATA SET

Data collection	Clustering method	
	PSO	KmeansPSO
Seeds	0.215	0.297
User Knowledge Modelling	0.083	0.209
Iris	0.205	0.392
Glass	0.214	0.386

The values of the silo divided by the number of clusters considered for each of the data sets are shown in the graphs below. As can be seen, in the clustering of all the siloet data sets for the proposed method, the number of clusters has better values.

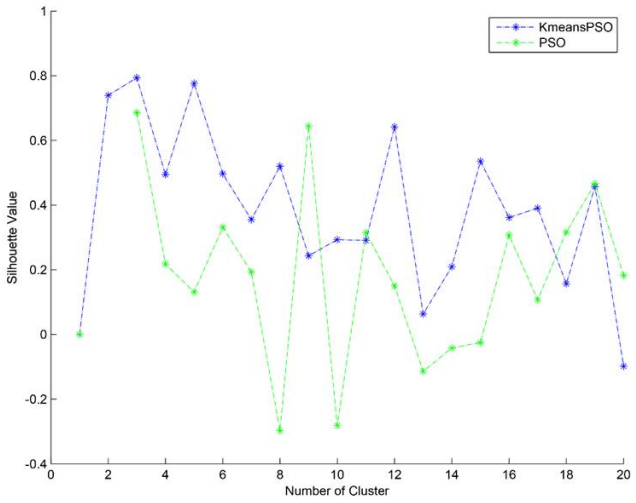

FIGURE I. THE SILOET CRITERION RESULTS OF GLASS DATASET CLUSTERING

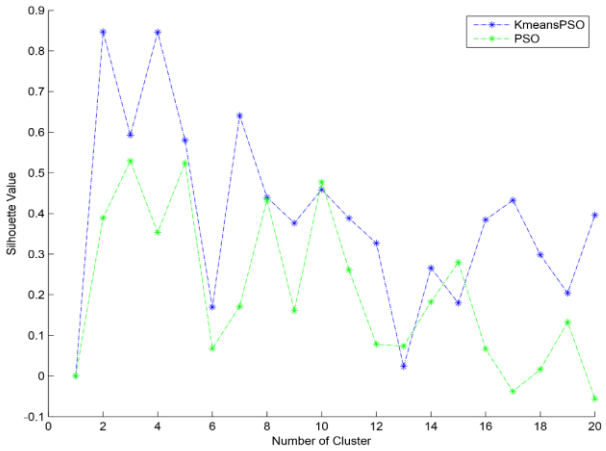

FIGURE II. THE SILOET CRITERION RESULTS OF IRIS DATASET CLUSTERING

In the proposed model, a new clustering method was created based on a combination of PSO and k-Means algorithms. As the two algorithms are shown in the diagrams and comparison results table, the proposed algorithm yields a better value than the silo criterion, which indicates an improvement in the status of the clustering algorithm in terms of sensitivity to the primary particles.

V. DATA MINING CHALLENGES IN HEALTHCARE

As we know that a lot of healthcare data is generated and stored by various healthcare organizations. But there are various challenges related to healthcare data which may play serious hurdles in the making proper decisions. The first challenge with healthcare data is the format of data being stored is different in different healthcare organizations. Till date there is no standard format is laid down for data being stored. In epidemic situations this

lack of standard format can make the epidemic situations even more badly. Suppose that an epidemic disease is spread within a country at its different geographical regions. The country health ministry requires that all the healthcare organizations must share their healthcare data with its centralized data warehouse for analysis in order to take all the essential steps so that epidemic situation may get resolve. But since the formats of data is different. Hence, the analysis of data may take longer time than usual. Due to this it may be possible that the situation may become out of control. The healthcare data is very useful in order to extract the meaningful information from it for improving the healthcare services for the patients. To do this quality of data is very important because we cannot extract the meaningful information from that data which have no quality. Hence, the quality of data is another very important challenge. The quality of data depends on various factors such as removal of noisy data, free from missing of data etc. All the necessary steps must be taken in order to maintain the quality in healthcare data.

Data sharing is another major challenge. Neither patients nor healthcare organizations are interested in sharing of their private data. Due to this the epidemic situations may get worse, planning to provide better treatments for a large population may not be possible, and difficulty in the detection of fraud and abuse in healthcare insurance companies etc. Another challenge is that in order to build the data warehouse where all the healthcare organizations within a country share their data is very costly and time consuming process.

VI. CONCLUSION AND FUTURE ISSUES

For effective utilization of data mining in health organizations there is a need of enhance and secure health data sharing among different parties. Some propriety limitations such as contractual relationships among researcher and health care organization are mandatory to overcome the security issues. There is also a need of standardized approach for constructing the data warehouse. In recent years due to enhancement of internet facility a huge datasets (text and non-text form) are also available on website. So, there is also an essential need of effective data mining techniques for analyzing this data to uncover hidden information.

VII. REFERENCES

[1] Hosseinkhani, Javad, et al. 2014, "Detecting Suspicion Information on the Web Using Crime Data Mining Techniques." International Journal of Advanced Computer Science and Information Technology 3(1) : 32-41.

[2] Hosseinkhani, J., S. Chuprat, and H. Taherdoost. 2012, "Criminal Network Mining by Web Structure and Content Mining." 11th WSEAS International Conference on Information Security and Privacy (ISP'12), Prague, Czech Republic September.

[3] Hosseinkhani, Javad, Suriayati Chuprat, and Hamed Taherdoost. 2012, "Discovering criminal networks by Web structure mining." Computing and Convergence Technology (ICCCT), 7th International Conference on. IEEE.

[4] Ghoreyshi, Saeedodin, and Javad Hosseinkhani. 2015, "Developing a Clustering Model based on K-means Algorithm in order to Creating Different Policies for Policyholders in Insurance Industry." International Journal of Advanced Computer Science and Information Technology (IJACSIT) 4(2): 46-53.

[5] Xueping Zhang, Jiayao Wang, Fang Wu, Zhongshan Fan and Xiaoqing Li, 2006, "A Novel Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and K-Medoids", Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications,Vol.1: 605 – 610.

[6] Thiemo Krink and Sandra Paterlini, 2006, "Differential Evolution and Particle Swarm optimization in Partitional Clustering", Journal of Computational Statistics and Data Analysis, Vol. 50(5): 1220-1247.

[7] Ian Davidson, Kiri L. Wagstaff, and Sugato Basu, 2006, "Measuring Constraint-Set Utility for Partitional Clustering Algorithms", Proceedings of the Tenth European Conference on Principles and Practice of Knowledge Discovery in Databases, Vol. 4213: 115-126.

[8] Taher Niknam and Babak Amiri, 2010, "An efficient hybrid approach based on PSO, ACO and K-Means for cluster analysis", Applied Soft Computing, Vol.10(1): 183-197.
[9] M. Arshad, 2012, "Implementation of Kea-Key phrase Extraction Algorithm by Using Bisecting K-Means Clustering Technique for Large and Dynamic Data Set", International Journal of Advanced Technology & Engineering Research, Vol.2(2), 134-137.
[10] Nor Ashidi Mat Isa, Amy A. Salamah and Umi Kalthum Ngah, 2009, "Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation", IEEE Transactions On Consumer Electronics, Vol.55(4): 2145-2153.
[11] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, 2008, "Top 10 algorithms in data mining", Knowledge and Information Systems, January, Volume 14(1): 1-37.
[12] O. A. Abbas, 2008, "comparisons between data clustering algorithms", international Arab journal of information technology, Vol. 5(3): 320-325.
[13] Y. S. Patil, M.B. Vaidya, 2012, "A Technical Survey on cluster analysis in data mining", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Vol 2(9): 03-513.
[14] C. McGregor, C. Christina and J. Andrew, 2012, "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). Vol. 765.
[15] R. Bellazzi and B. Zupan, 2008, "Predictive data mining in clinical medicine: current issues and guidelines", Int. J. Med. Inform., Vol. 77 (2008): 81-97.
[16] M. Kumari and S. Godara, 2011, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", International Journal of Computer Science and Technology (IJCST) ISSN: 2229- 4333, Vol. 2(2): 304-308.
[17] S. Gupta, D. Kumar and A. Sharma, 2011, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering, Vol. 2(2):188-195.
[18] T. S. Chen, T. H. Tsai, Y. T. Chen, C. C. Lin, R. C. Chen, S. Y. Li and H. Y. Chen, 2005, "A Combined K-Means and Hierarchical Clustering Method for improving the Clustering Efficiency of Microarray", Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems.
[19] Chipman and R. Tibshirani, 2006, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics, Vol. 7(2): 286-301.
[20] J. J. Tapia, E. Morett and E. E. Vallejo, 2009, "A Clustering Genetic Algorithm for Genomic Data Mining", Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, Vol. 204: 249-275.
[21] T. H. A. Soliman, A. A. Sewissy and H. A. Latif, 2010, "A Gene Selection Approach for Classifying Diseases Based on Microarray Datasets", 2nd International Conference on Computer Technology and Development (lCCTD 2010).
[22] S. Belciug, 2009, "Patients length of stay grouping using the hierarchical clustering algorithm", Annals of University of Craiova, Math. Comp. Sci. Ser., ISSN: 1223-6934, Vol. 36(2): 79-84.
[23] Schulam. P., Wigley. F., Saria. S. 2015, "Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2956-2964.