2021 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-7281-9766-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE50685.2021.9427598

Clustering Application for Data-Driven Prediction of Health Insurance Premiums for People of Different Ages

Tallal Omar
Computer Science and Engineering
Oakland University
Rochester, Michigan, USA
tallalomar@oakland.edu

Mohamed Zohdy
Electrical and Computer Engineering
Oakland University
Rochester, Michigan, USA
zohdyma@oakland.edu

Julian Rrushi
Computer Science and Engineering
Oakland University
Rochester, Michigan, USA
rrushi@oakland.edu

***Abstract*— A health insurance premium is a monthly fee that is paid in a health plan to typically pay for medical, surgical, prescription drug and sometimes dental expenses incurred by the insured. Since 2010, the affordable care act has prohibited insurance companies from denying coverage to patients with pre-existing conditions and has allowed children to remain on their parents' insurance plans until they reached the age of 26. Creating the policy is a really important and challenging task. In order to determine health insurance premium quotes, there are several factors that have to be taken into consideration when defining a premium, such as pre-existing diseases, age, gender, family medical history, lifestyle, etc. In this paper, a combination of the K-means algorithm and the Elbow method is developed to accurately group people in an optimal number of clusters based on similarity. Based on this evaluation, the health insurance premium quote via the provided factors to predict the range of the health insurance premium quote for each group of people.**

Keywords — k-means technique, Elbow technique, clustering technique, data mining, health insurance.

I. INTRODUCTION

Many approaches have been proposed to accelerate the large scale data [2]. In the last few decades, there are some clustering approaches that have been introduced for the purpose of better performance in several applications [3]. Previous research has resulted in a number of different algorithms for rule discovery [5]. The clustering techniques might be split into several methods [6 - 9], namely situation awareness for intelligent online learning platforms [7], Parkinson's disease [8], hybrid technique k-means and artificial neural networks [10-11]. The k-means clustering algorithm generates k points as initial centroids arbitrarily based on their similarity to each other [12]. The circular k-means (CK-means) clusters vectors containing directional information [13]. A research conducted by Máximo et al. described a new competitive k-means algorithm to address the inconsistent results of traditional k-means, which scales poorly for large data sets [2]. Kumar conducted research to present a taxonomy of clustering techniques [3]. Jasser et al. explored combining machine learning and intelligent systems with k-means [7]. Marutho et al. researched the integration of traditional k-means with the Elbow method to determine the optimal number of clusters in the k-means algorithm. The authors applied their research to news headline data [9]. Yang et al. discussed a nonlinear function that integrates a dimensionality reduction (DR) and k-means partitioning to cluster for latent data representations [14].

An alternative scheme to k-means is the Elbow method. It is commonly used to identify the optimal number of clusters in a data set. The method runs k-means clustering on any number of data sets for a range of values for k-clusters. Syakur et al. explored the combination of the traditional k-means algorithm with the Elbow method to identify the optimal numbers of clusters [16]. Purnima et al. proposed an energy enhanced protocol that uses k-means clustering to select the cluster heads for each cluster by combining it with the Elbow method [17]. Huang improved the k-means algorithm by also combining it with the Elbow method [18].

The goal of this research is to introduce a new clustering technique, which integrates the traditional k-means algorithm with three other methods. Firstly, the normalization/standardization technique was applied to scale the data set. Secondly, the principal component analyses (PCA) were applied to reduce the data set. Thirdly, the Elbow method was used to define the parameter k, which, again, is the optimal number of clusters. Fourth, the k-means algorithm

was applied to provide the optimal number of clusters, k, as input to the algorithm on a health insurance cross sell prediction data set.

This research considered a sample size of 381,110 people using 11 features, as summarized in Table 2. Annual health premiums are based on a variety of factors, including age, pre-existing conditions, gender, family medical history, and lifestyle. Then, the program generates a quote based on age. In general, the lower the age, the lower the premium.

The center for disease control (CDC) announced that in each year in the United States, about three million people are non-fatally injured in motor vehicle crashes, and that crash related injuries are very costly. Medical care costs and productivity losses associated with injuries and deaths from motor vehicle crashes exceeded $75 billion in 2017. Therefore, this gives us the motivation to work on the data set provided by health insurance companies for cross sell prediction. We build a model that helps insurance companies predict new health premium quotes that lower health insurance premiums for older adults based on new clustering factors.

The rest of this paper is organized as follows. In Section II we describe a new methodology and its principles of data analysis as applied in this research. In Section III we present the results of this new methodology on real-world data sets. Section IV summarizes our contribution, proposes future work, and concludes the paper.

II. METHODOLOGY

In our proposed methodology, we define the optimal number of clusters through a process of evaluation that consists of five steps. Data sets for 381,110 people were passed through five steps. We start with a preprocessing step, which scales and standardizes the data set for different parameter ranges. For example, the annual premium field came in five digits, and the age field came in two digits. These large data might dominate the objective function and make the estimator unable to learn from other features as correctly as expected [21].

The second step is PCA [22], which is aimed at reducing the data set dimension. The third step is the Elbow method [17], which we use to define the optimal number of k-clusters. The forth step is the k-means clustering algorithm [10-11], which clusters and groups people based on the similarity of their parameters.

The final step consists of calculating and recording the Error Sum of Squares (SSE), which is the sum of the squared differences between each independent variable in the data set with its group's mean. This helps with determining the number of clusters.

The overall methodology is illustrated by the flowchart of Figure 1.

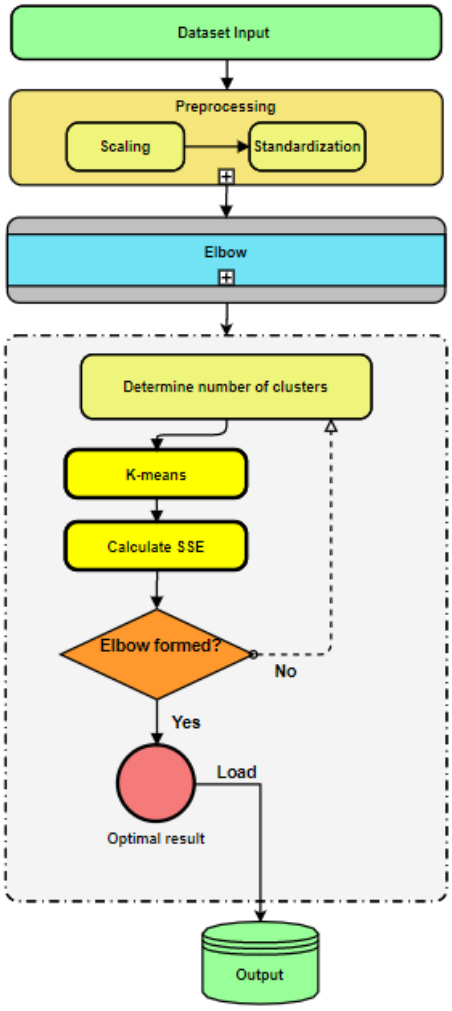

Figure 1. *The process flowchart for the proposed data analysis system.*

A. Preprocessing

Scaling is a common requirement for many machine learning estimators for any given data sets. This is known as the method of standardizing and normalizing, and is implemented as a Scikit-Learn function in the Python programming language. We

take particular care of this important detail, since the estimator might behave badly if the data set is not scaled and standardized for each individual feature. Features need to be processed with mean of zero and no unit variance.

In mathematical terms, we distribute and centralize the data set by removing the mean value of each individual feature. We then scale it by dividing non-constant features by their standard deviation. In this process, the standard deviation is represented as follows:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

In equation (1), z represents the standardized and normalized value; x is the individual feature value of the data point; μ is the population mean; and σ is the population standard divisor for the data set.

B. *Principal Component Analysis.*

PCA is used in this work as a dimensionality reduction method. By definition, PCA reduces the dimensionality of large data sets by transforming a large data set of variables into a small data set. Here, PCA preserves as much information as possible in that it is constructed as linear combinations or mixtures of the initial variables. One of the most important applications of PCA is the speeding up of machine learning algorithms [22]. Using the health insurance cross sell prediction data set is practical, since this data set has 381,110 rows and 11 feature columns. This step has to be done after scaling the data set.

C. *K-Means Method.*

The k-means clustering algorithm is a simple method for partitioning *n* data points in *k* groups of clusters. K-means was first established by J. B. MacQueen [23]. This clustering approach has been applied for cluster analyses in data mining and pattern recognition to minimize within-cluster variances, i.e. squared Euclidean distances. Furthermore, k-means has been defined as one of the simplest data mining partitioning and clustering techniques that implement the Euclidean distance function. The goal of the k-means algorithm in this work is to reduce the error sum of squares along with the error criterion. Both of these factors are the backbone of discovery of the optimal value of *k* divisions that meet specific criteria. In this work, these are the functional objectives of k-means, even for some non-numeric data set. There are several other advantages to the k-means clustering approach, namely brevity, efficiency, and swiftness [24].

The k-means method relies on initial data points, which are foggy and randomly partitioned. Selecting initial samples of a random centroid in a cluster usually is directed at various outcomes [25]. The k-means method relies on the gradient method, which is an optimization approach that constantly updates parameters to get the optimal peak value. The native algorithm is illustrated via the Hartigan-Wong algorithm [26]. To define the total within-cluster variation as the sum of the squared distances between values and its corresponding centroid, we present the following:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (2)$$

In equation (2), $W(C_k)$ represents the total within-cluster variation; x_i is a data point for a cluster; C_k indicates a cluster for each data point; and μ_k is the mean value of the data points that is assigned to cluster C_k. We define TW as the sum of total within-cluster variation, as follows:

$$TW = \sum_{K=1}^{K} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (3)$$

D. *Elbow Method*

In this work, the Elbow technique is used for cluster analyses to determine the optimal number of clusters in k-means. It consists of plotting the variation, and hence the function would provide the optimal number of clusters, selecting the elbow of the curve as the number of clusters to use. Adding more clusters does not provide any better modeling of the data set. The first clusters insert large amounts of details, but at a certain point, the marginal number of clusters attained falls significantly and thus provides an angle in the graph [27]. The optimal number of *k*, which, again, is the number of clusters selected, is referred to as the Elbow criterion.

The method will start with *k=2*, and then will increase *k* by one every time. The value of *k* calculates the sum of squared errors (SSE). Then, we plot a line chart of the SSE for each value of *k*. When it reaches to the point where the line chart looks like an arm, then the elbow is performed and the value of *k* becomes the optimal number of clusters we could get for k-means, as in [28]. The Elbow method is described in equation (4) as the within-groups sum of squares (WSS), where the squared average distance

of all the data points for a cluster is a distance that is statistically measured from the group means to the same cluster centroid [23], as shown below:

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2 \qquad (4)$$

By combining the k-means and Elbow methods, we can predict the optimal number clusters *k*. The Elbow method can be expressed by the sum of the squared error [24], as follows:

$$SSE = \sum_{k-1}^{k} \sum_{xi \in Sk} \|x_i - c_k\|_2^2 \qquad (5)$$

The SSE equals zero when *k* becomes equal to the number of data points in the data set. Then, each data point is its own cluster, and thus there is no error between it and the center of its cluster. Here, k is equal to many clusters that formed C, which is the i^{th} cluster, with X, representing the total data points in each cluster.

III. Evaluation Results

As we discussed earlier in this paper, we tested our approach on a data set that was provided by Health Insurance Cross Sell Prediction [1]. Our approach groups people into clusters based on their similarity via different factors. Table 1 summarizes a sample of the data set. We are using acronyms instead of full denominations to make them fit in the table. Gr stands for Gender, DL for Driving License, RC for Region Code, PI for Previously Insured, VA for Vehicle Age, VD for Vehicle Damage, AP for Annual Premium, V for Vintage, and finally R for Response.

TABLE I. Sample raw data set for people of different ages

id	Gr	Age	DL	RC	PI	VA	VD	AP	V	R
1	1	44	1	28	0	2	1	40454	217	1
2	1	76	1	3	0	1	0	33536	183	0
3	1	47	1	28	0	2	1	38294	27	1
4	1	21	1	11	1	0	0	28619	203	0
5	0	29	1	41	1	0	0	27496	39	0
6	0	17	0	33	0	0	1	2630	176	0

As illustrated in Figure 1, we start with a preprocessing stage. Scaling the data set is necessary, since variance changes are frequent. For example, in the data set at hand, number 1 represents true, number 0 represents false, age is comprised of 2 digits, and the annual premium is made of up to 5 digest, which is quite large in terms of magnitude.

Let us assume for a moment that the data set has two features, namely age and weight. Clearly these features are not directly comparable. For example, one year is not equivalent to one pound, and may not have the same level of importance in sorting. In a situation where one feature has a much greater range of possible values than others in terms of magnitude, it may end up being the primary driver of what defines clusters.

The feature with the wider range of values has greater distances between values. Scaling helps to make the relative weight of each variable equal by converting each variable to a unitless measure of relative distance. Table 2 shows how PCA is calculated and reduces data dimensions from eleven features to only two features, while preserving as much information as possible.

TABLE II. PCA calculated after scaling data

principal component 1	principal component 2
3.306373	0.330084
1.511856	-1.810449
3.203895	-0.210509
-1.748110	-0.672592
-1.918238	-0.145919
-1.507125	3.158955
-0.566121	1.480411
2.197677	0.379354
-2.360887	-1.290697

Table 3 presents SSE as calculated before and after scaling the data set.

TABLE III. SSE calculated before and after scaling

SSE before scaling	SSE after scaling
4725741986646696.0	4573308.0
1266129313088534.0	3703545.6
625461297499750.9	3021959.6
401233698043987.0	2869976.3
297444697731995.75	2690037.4
241057110514856.75	2480240.3

207069601474127.47	2369331.5
184999154227273.72	2263327.1
169887826023748.3	2189123.0

We can see in Table 3 that the score for SSE decreases after scaling at each iteration. Our goal is to minimize SSE, and thus we choose a small value of *k* that still has a low score for SSE.

In Figure 2, the elbow method has been implemented on the data set before scaling. The elbow curve is at *k* equal to 2. The estimator behaved badly, and it is hard to determine the optimal number of clusters.

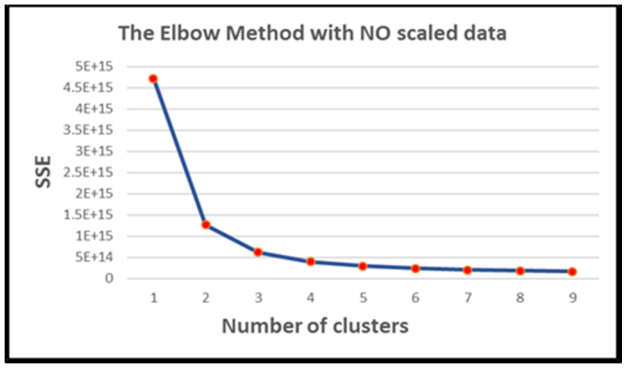

Figure. 2. The Elbow method before scaling data

In Figure 3, the elbow method has been implemented on the data set after scaling. In this case, we can clearly see the elbow curve at k equal to 3. The estimator behaved as expected.

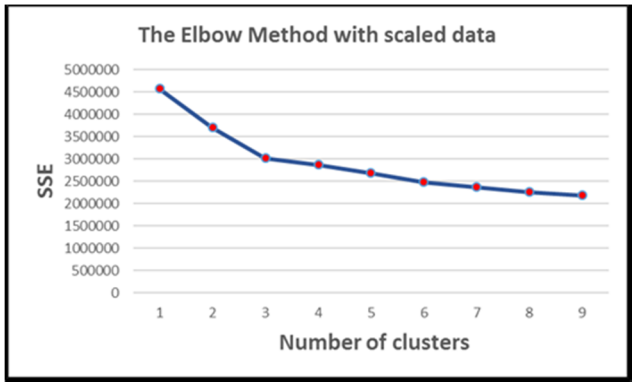

Figure. 3. The Elbow method after scaling the data

In Figure 2, the Elbow method was at *k* equal to 2 because the data set features contained different ranges of values. Consequently, it is important to scale the values of the features to the same range to get more accurate results from the Elbow method. The difference in the elbow between scaled and unscaled data sets is evidently visible.

IV. Conclusions and Future Work

This paper provided a simple and efficient prediction model for insurance companies to help them predict and estimate health insurance premium quotes. The k-means algorithm was combined with the Elbow method to obtain the optimal number of clusters. Two scenarios were considered for the Elbow method to show the importance of scaling the data before clustering. For future work, more advanced data filtering could be researched. Furthermore, k-means clustering, self-organizing feature maps, or other clustering methods, can be explored further for more optimized clustering.

Acknowledgment

The authors would like to thank the Health Insurance Cross Sell Prediction team for providing the data set used in this work.

References

[1] Kumar, A. (2020, August). Health Insurance Cross Sell Prediction, Version 1. Retrieved September 10, 2020 from https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction.

[2] R. M. Esteves, T. Hacker and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Data sets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, 2013, pp. 17-24, doi: 10.1109/CloudCom.2013.89.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn. “Data clustering: a review”, ACM Comput. Surv.1999.

[4] Linghong Zou and Luling Zhou, "The Discussion about preservation and increasing the value of Social health insurance fund," 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Dengleng, 2011, pp. 4824-4827, doi: 10.1109/AIMSEC.2011.6011040.

[5] Reynolds, A.P., Richards, G., de la Iglesia, B. et al. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. J Math Model Algor 5, 475–504 (2006). https://doi.org/10.1007/s10852-005-9022-1

[6] S. Singh and A. Mishra, "Clustering analysis for large scale data sets," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 1-4, doi: 10.1109/CCAA.2015.7148353.

[7] Jasser, J., Ming, H., & Zohdy, M. A. (2017, July). Situation-Awareness in Action: An Intelligent Online Learning Platform

(IOLP). In International Conference on Human-Computer Interaction (pp. 319-330). Springer, Cham.

[8] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L., & Dauer, W. (2018). Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. Scientific reports, 8(1), 1-21.

[9] Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 533-538). IEEE.

[10] Sharma, M., Purohit, G. N., & Mukherjee, S. (2018). Information retrieves from brain MRI images for tumor detection using hybrid technique K-means and artificial neural network (KMANN). In Networking communication and data knowledge engineering (pp. 145-157). Springer, Singapore.

[11] Kumar, J., & Vashistha, R. (2017, February). Estimation of inter-centroid distance quality in data clustering problem using hybridized K-means algorithm. In 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-7). IEEE.

[12] Yedla, M., Rao, S.S., Pathakota, & Srinivasa, T. (2010). Enhancing K-means Clustering Algorithm with Improved Initial Center.

[13] D. Charalampidis, "A modified k-means algorithm for circular invariant clustering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1856-1865, Dec. 2005, doi: 10.1109/TPAMI.2005.230.

[14] Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017, August). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3861-3870). JMLR. org.

[15] Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 533-538). IEEE.

[16] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration K-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering (Vol. 336, No. 1, p. 012017). IOP Publishing.

[17] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

[18] Huang, G. X., & Lin, D. (2019, October). Clustering Analysis and Visualization of Terrorist Attack Data. In Proceedings of the 2019 International Conference on Video, Signal and Image Processing (pp. 136-136).

[19] Tefft, B.C. (2017). Rates of Motor Vehicle Crashes, Injuries and Deaths in Relation to Driver Age, United States, 2014-2015. AAA Foundation for Traffic Safety.

[20] Global status report on road safety 2018. Geneva: World Health Organization; 2018. Licence: CC BYNC-SA 3.0 IGO

[21] Tang, J., Wang, D., Zhang, Z., He, L., Xin, J., & Xu, Y. (2017). Weed identification based on K-means feature learning combined with convolutional neural network. Computers and Electronics in Agriculture, 135, 63-70.

[22] A. Izzuddin, "Optimasi Cluster pada Algoritma KMeans dengan Reduksi Dimensi Data set Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen," Energy J. Ilm. Ilmu-Ilmu Tek., vol. 5, no. 2, pp. 41–46, 2015.

[23] MacQueen, J. B. (1965). On the asymptotic behavior of K-means (No. WMSI WORKING PAPER-89). CALIFORNIA UNIV LOS ANGELES WESTERN MANAGEMENT SCIENCE INST.

[24] Jamadi, N. A., Siraj, M. M., Din, M. M., Mammy, H. K., & Ithnin, N. (2018). Privacy Preserving Data Mining Based on Geometrical Data Transformation Method (GDTM) and K-Means Clustering Algorithm. International Journal of Innovative Computing, 8(2).

[25] Q. Qiu, Q. Zhang and K. Guo, "Grey Kmeans algorithm and its application to the analysis of regional competitive ability," 2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference, Chongqing, 2014, pp. 249-253.

[26] Amaral, G. J., Dore, L. H., Lessa, R. P., & Stosic, B. (2010). K-means algorithm in statistical shape analysis. Communications in Statistics—Simulation and Computation®, 39(5), 1016-1026.

[27] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

[28] Nguyen, Hoang, et al. "A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and cubist algorithms." Applied Soft Computing 77 (2019): 376-386.