ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20180720122815
Original scientific paper

Auto Insurance Business Analytics Approach for Customer Segmentation Using Multiple Mixed-Type Data Clustering Algorithms

Kai ZHUANG, Sen WU, Xiaonan GAO

Abstract: Customer segmentation is critical for auto insurance companies to gain competitive advantage by mining useful customer related information. While some efforts have been made for customer segmentation to support auto insurance decision making, their customer segmentation results tend to be affected by the characteristics of the algorithm used and lack multiple validation from multiple algorithms. To this end, we propose an auto insurance business analytics approach that segments customers by using three mixed-type data clustering algorithms including k-prototypes, improved k-prototypes and similarity-based agglomerative clustering. The customer segmentation results of these algorithms can complement and reinforce each other and demonstrate as much information as possible to support decision-making. To confirm its practical value, the proposed approach extracts seven rules for an auto insurance company that may support the company to make customer related decisions and develop insurance products.

Keywords: auto insurance; business analytics approach; clustering; customer segmentation; mixed-type data

1 INTRODUCTION

Insurance companies are indispensable to people's life, provide appropriate services and improve people's welfare. Since customer is considered as a critical factor of insurance companies in producing revenue and improving profitability, how to obtain and keep customers is the major problem in insurance companies [1]. The profitability of insurance companies mainly depends on the services they offer and on meeting the customer demand on a regular basis, so a good customer related strategy must be found to analyse customer's features and biases. Customer segmentation is a powerful tool to divide customers into different clusters and analyse their characteristics. Auto insurance is one of the important components in insurance industry, which is profitable and lucrative. In this paper, we focus on the research of customer segmentation in auto insurance companies.

Currently, most of the existing researches about auto insurance companies focus on fraud detection [2-4], premium calculation [5], feature selection [6] and customer segmentation [7]. The first three research objectives only support auto insurance companies to accomplish daily works but have not utilized the mass customer related data to segment customer and discover helpful information for companies. A few existing researches on ordinary insurance customer segmentation only utilize one algorithm to analyse customer related data. Fuzzy Analytic Network Process (FANP) based weighted RFM (Recency, Frequency, Monetary value) model [8] combines the k-means paradigm to learn Gidden knowledge and information by segmenting customers of auto insurance. Auto Insurance Customers Segmentation Intelligent Tool [9] that is a two-phase model segments customers of auto insurance companies based on risk. These methods used for customer segmentation merely exploit one algorithm to do analyses which incline to produce defective analysis results impacted by the characteristics of the exploited one algorithm. We desire to mine information reflecting real customers rather than being impacted by the used algorithm. Therefore, we use multiple algorithms to segment customers in auto insurance companies and analyse the characteristics of different categories of customers to provide some customer related strategies for decision makers. In this way, the multiple analysis results can complement and reinforce each other and demonstrate as much information as possible to assist in the decision making of auto insurance companies.

Obtaining labels is costly and time consuming in the auto insurance industry, and much unlabelled customer related data is ready for analysis, therefore, we analyse the customer related data of auto insurance companies by using unsupervised data mining technologies. Clustering algorithms are typical unsupervised learning technologies. Additionally, in real-world tasks, customer related data contains both numerical (e.g. new car purchase price and vehicle age) and categorical (e.g. policy nature and policy state) attributes simultaneously. Hence, we select several different mixed-type data clustering algorithms to segment customers and analyse characteristics.

The purpose of this paper is to segment customers of auto insurance companies using multiple mixed-type data clustering algorithms and analyse characteristics of different customers. To this end, more convincing and accurate analysis results could be obtained which are not affected by clustering algorithms to support decision making.

The main contributions of this paper can be summarized as follows:

(1) An auto insurance business analytics approach is proposed, in which multiple mixed-type data clustering algorithms are utilized to segment customers. Since the mixed-type data clustering is a challenging task, there is no perfect method to deal with this problem. Compared with processes that only use one algorithm, the results of different algorithms can complement and reinforce each other, and we could extract more convincing and accurate information to assist in the auto insurance decision making.

(2) The practical value of the approach is validated. We exploit a real case to demonstrate our approach to extract appropriate rules for the auto insurance company. The auto insurance company can develop more appropriate insurance products for different customers based on the extracted rules to keep and attract customers.

2 MIXED-TYPE DATA CLUSTERING ALGORITHMS

We review a few typical techniques for clustering mixed-type data. Currently, the clustering algorithms for processing mixed-type data can be classified into three

types [10, 11]: converting numerical attributes into categorical attributes, converting categorical attributes into numerical attributes, and clustering mixed-type data directly. The first two clustering algorithms may lead to the information loss and impact the accuracy of results [12]. Therefore, we focus on the third type of mixed-type data clustering algorithm.

K-prototypes [13] is a classical mixed-type data clustering algorithm by combining k-means for numerical attribute data and k-modes for categorical attribute data, which is very cost-effective. Similarity-Based Agglomerative Clustering (SBAC) [14] based on Goodall similarity metric [15] was proposed which circumvents parameter determination. Improved k-prototypes [16] can cluster incomplete mixed-type data directly and eliminate the sensitivity of initial prototypes. Evidence-Based Spectral Clustering algorithm [17] integrates the spectral clustering frame and evidence-based similarity computation method to cluster mixed-type data. Moreover, some similarity or dissimilarity of mixed-type data was proposed. A dissimilarity for mixed-type data derived from the probabilistic model was proposed in [18]. And a unified similarity coefficient [19] was presented based on the importance of the categorical attribute values. A similarity measure [20] between any two mixed-type data objects was proposed based on the uncertainty of the attribute values.

Among the aforementioned algorithms, k-prototypes is very efficient, but the parameter determination is complicated, and its clustering result is sensitive to the initial prototypes. Improved k-prototypes could reduce the sensitivity of initial prototypes by determining prototypes based on neighbours. Additionally, for SBAC, there is no need to pre-set parameters and it clusters mixed-type data very effectively, but the time cost of computation is high. Since these three algorithms could complement each other, we select them to construct the business analytics approach for customer segmentation.

Next, we briefly review these three mixed-type data clustering algorithms. Let x_i be a data object described by m attributes, m_n and m_c are the number of numeric and categorical attributes respectively, $m_n + m_c = m$, x_{il}^n represents the l^{th} numeric attribute value of X_i and x_{il}^c represents the l^{th} categorical attribute value of X_i.

2.1 K-prototypes

K-prototypes eliminate the numeric data limitation of k-means, but preserve the efficiency by inheriting its paradigm. A dissimilarity measure for mixed-type data is proposed by combining k-means and k-modes. It can be computed as follows:

$$d\left(X_i, X_j\right) = \sum_{l=1}^{m_n}\left(x_{il}^n - x_{jl}^n\right)^2 + \gamma\sum_{l=1}^{m_c}\delta\left(x_{il}^c, x_{jl}^c\right) \tag{1}$$

$$\delta\left(p,q\right) = \begin{cases} 0, p = q \\ 1, p \neq q \end{cases} \tag{2}$$

where γ is a weight for categorical attributes that is usually set to the ratio of the number of categorical attributes to the number of all attributes.

The steps of k-prototypes are the same as those of k-means. Firstly, pre-set the number of clusters k and initialize k-prototypes. Secondly, allocate each data object into the cluster in which the dissimilarity between this data object and cluster prototype is minimum. Thirdly, update prototypes and reallocate data objects until no prototype can be updated.

2.2 Improved K-prototypes

Improved k-prototypes refines the k-prototypes from two aspects. It not only can cluster incomplete data with no need to impute the missing values, but also can avoid sensitivity of initial prototypes.

Improved k-prototypes defines a new dissimilarity computing method called 'Incomplete Set Mixed Dissimilarity (ISMD)', that computes dissimilarity between two incomplete mixed-type data objects with no need for imputing missing values in advance to avoid an estimation that may cause error. The categorical and numerical attribute dissimilarity between X_i and X_j is defined respectively as follows:

$$\delta_k\left(X_i, X_j\right) = \begin{cases} 1, x_i^k \neq x_j^k \wedge x_i^k \neq \text{"*"} \wedge x_j^k \neq \text{"*"} \\ 0, x_i^k = x_j^k \vee x_i^k = \text{"*"} \vee x_j^k = \text{"*"} \end{cases} \tag{3}$$

$$d_l\left(X_i, X_j\right) = \begin{cases} \dfrac{\left|x_i^l - x_j^l\right|}{Max_l - Min_l}, x_i^l \neq \text{"*"} \wedge x_j^l \neq \text{"*"} \\ 0, x_i^l = \text{"*"} \vee x_j^l = \text{"*"} \end{cases} \tag{4}$$

where $\delta_k(X_i, X_j)$ and $d_l(X_i, X_j)$ respectively represent dissimilarity between X_i and X_j in categorical attribute k and numerical attribute l. "*" represents the missing values.

In addition, improved k-prototypes initializes k prototypes according to neighbours, the k (number of clusters) data objects with the most neighbours are selected as initial prototypes. This method reduces the randomness of the clustering result.

The steps of improved k-prototypes are similar to those of k-prototypes and can be defined in four steps. Firstly, initialize the prototypes based on neighbours. Secondly, allocate each data object into the cluster in which the dissimilarity between this data object and cluster prototype is minimum. Thirdly, update prototypes and reallocate data objects until there is no prototype that can be updated. Finally, fill the missing values based on clustering result.

Improved k-prototypes not only inherits the effectiveness of k-prototypes, but also eliminates the sensitivity of initial prototypes and computes the dissimilarity of incomplete mixed-type data more accurately.

2.3 Similarity-Based Agglomerative Clustering

Similarity-Based Agglomerative Clustering (SBAC) utilizes a similarity measure proposed by Goodall [15] that has no need for pre-setting parameters in advance and constructs a dendrogram to extract the final clustering result heuristically.

The similarity CS_{ij}^{l} between two data objects X_i and X_j in l^{th} categorical attribute is computed as follows:

$$CS_{ij}^{l} = 1 - \sum_{k \in MSFVS\left(x_{il}^{c}, x_{jl}^{c}\right)} \left(p_k\right)_l^2 \tag{5}$$

where $(p_k)_l$ is the probability of occurrence of value x_{kl}^{c} in the data set, $MSFVS\left(x_{il}^{c}, x_{jl}^{c}\right)$ is the set of all pairs of values for l^{th} categorical attribute that are equally or more similar to the pair $\left(x_{il}^{c}, x_{jl}^{c}\right)$.

Analogously, the similarity NS_{ij}^{l} between two data objects X_i and X_j in l^{th} numerical attribute is computed as follows:

$$NS_{ij}^{l} = 1 - \sum_{k,m \in MSFSS\left(x_{il}^{n}, x_{ij}^{n}\right)} \left(p_k\right)_l \left(p_m\right)_l \tag{6}$$

where $(p_k)_l$ and $(p_m)_l$ are the probabilities of occurrence of x_{kl}^{n} and x_{ml}^{n}, $MSFSS\left(x_{il}^{n}, x_{jl}^{n}\right)$ is the set of all pairs of values for l^{th} numerical attribute that are equally or more similar to the pair $\left(x_{il}^{n}, x_{jl}^{n}\right)$.

Next, the similarity S_{ij} between X_i and X_j in all attributes can be computed as follows:

$$g_{ij}^{2} = 2\sum_{l=1}^{m_c}\left(1 - \frac{CD_{ij}^{l}\ln CD_{ij}^{l} - \left(CD_{ij}^{l}\right)'\ln\left(CD_{ij}^{l}\right)'}{CD_{ij}^{l} - \left(CD_{ij}^{l}\right)'}\right) - 2\sum_{l=1}^{m_n}\ln ND_{ij}^{l} \tag{7}$$

$$S_{ij} = 1 - e^{-\frac{g_{ij}^{2}}{2}} \times \sum_{l=0}^{m_c+m_n-1}\frac{\left(\frac{1}{2}g_{ij}^{2}\right)^{l}}{l!} \tag{8}$$

where CD_{ij}^{l} and ND_{ij}^{l} are respectively equal to 1 minus CS_{ij}^{l} and NS_{ij}^{l}. Using this similarity measure, the dissimilarity matrix among data set can be calculated.

Based on the dissimilarity matrix, we can implement SBAC by three steps. Firstly, regard each data object as a cluster and merge the two clusters with the minimum dissimilarity in the matrix into one cluster. Secondly, delete the rows and columns corresponding to the two merged clusters and update the dissimilarity matrix by computing the dissimilarity between newly merged cluster and each other existing cluster. Finally, repeat the first two steps until all data objects are merged into the same cluster.

After conducting the above steps, a dendrogram is constructed from bottom to top. Next, we need to extract appropriate parts from the dendrogram to construct final clustering result. A threshold t (it is multiple of dissimilarity of root node) is given for depth-first traversal on the dendrogram. When the D-value between the dissimilarity of former node and current node is larger than t, the cluster represented by current node can be selected as one cluster in the clustering result. And then return to former node for continuous traversal until there is no node for traversal. Finally, the clustering result can be obtained.

3 BUSINESS ANALYTICS APPROACH FOR CUSTOMER SEGMENTATION

We propose a business analytics approach for customer segmentation which exploits three clustering algorithms to segment customers and recognize the characteristics of different customers, by analysing the categories of customers who have purchased auto insurance products.

The procedure of business analytics approach shown in Fig.1 is divided into three phases: (a) data collection and preparation, (b) customer segmentation and (c) customer characterization and integration. Next, we summarize the phases of the approach.

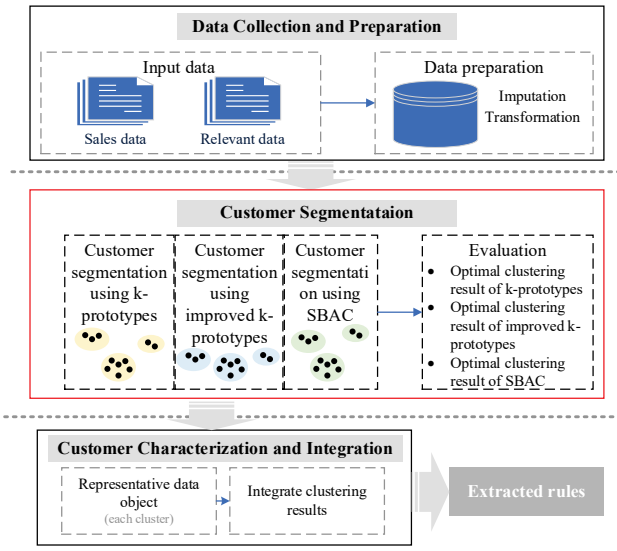

Figure 1 Business analytics approach for customer segmentation

3.1 Data Collection and Preparation

For auto insurance companies, sales data is the core data set for data segmentation, which contains products, turnover, purchase time, customer information and so on. It could reflect the essential information of an auto insurance transaction. Moreover, if more relevant data of other types could be obtained, that is a good way to expand studying data besides sales data. In this approach, sales data is indispensable, and other relevant data is encouraged to perfect studying data.

In reality, the sales data and other relevant data of auto insurance companies are usually incomplete and mixed-type. Thus, data preparation is required, which includes data imputation for imputing missing data and data transformation for digitizing categorical attribute values.

3.2 Customer Segmentation

This phase is essential for the approach that performs the following four tasks: customer segmentation using k-prototypes, improved k-prototypes and SBAC, evaluation, to ensure that the convincing and accurate results can be obtained.

First, use k-prototypes algorithm to segment customers, the number of clusters k should be pre-set. To ensure accurate customer segmentation and analysis, the optimal clustering result is always expected. k is given according to the data size and the characteristics of auto insurance companies. If there is no experience for determining k, we suggest that a wider range of k could be set, and then the optimal clustering result would be selected by the following evaluation step.

Next, use improved k-prototypes to segment customers, the number of clusters k also needs to be pre-set. Similarly, k is given according to the data size and the characteristics of auto insurance companies and could be set with a large range if there is no idea to pre-set it.

For SBAC, if we want to gain the final clustering result, we have to extract the appropriate parts from the dendrogram to construct the clustering result. The threshold t is used to achieve extraction and needs to be pre-set in advance. The determination of t depends on the dendrogram constructed and the dissimilarity of each node. We encourage to set a wider range of t for selecting the optimal clustering result.

After all clustering results are prepared, we need to select the optimal clustering result of each algorithm. Since the labels of the real auto insurance data are unknown, an internal cluster validation metric should be selected to find the optimal clustering result.

3.3 Customer Characterization and Integration

Here, the optimal clustering results are extracted, and next we need to identify the final customer segmentation and analyse the characteristic of each segment to mine new information and knowledge for decision supporting. We suggest computing the representative data object of each cluster. The mode values of categorical attributes and the mean values of numerical attributes are calculated as representative attribute values of each cluster. Furthermore, the most important attribute that directly reflects the profitability should be selected and guides companies or researchers to identify property of cluster.

The characteristics of different clusters can be discovered by analysing the representative data object of each cluster. Then we integrate the clustering results and extract customer related rules for decision supporting. We assume that the characteristics that more than two clustering results are consistent should be noted and analysed. The corresponding rules would be extracted after more discussion. For example, if the clustering results of k-prototypes and SBAC show that new car purchase price is always the highest when accumulative paid-in amount is the highest, and after analysis, this is reasonable, we could extract the rule that in the cluster corresponding to the highest accumulative paid-in amount, new car purchase price is always the highest. In this way, auto insurance company should focus on the customers who own luxury cars. On this basis, auto insurance company could develop more appropriate customer related strategies to keep and attract customers.

Table 1 Table of valuable attributes

Valuable attributes			
Policy nature	Policy state	Vehicle type	Seating capacity
License plate color	Document type	Insurance type	New car purchase price
Vehicle age	Accumulative paid-in amount	Written premium	Tonnage

Table 2 Comparison table of categorical attributes digitization

	0	1	2	3
Policy nature	New insurance	Renewal of insurance	--	--
Policy state	Effective	Insurance cancellation	--	--
Vehicle type	Passenger car with 6 or less seats	Passenger car with more than 6 and less than 10 seats	Passenger car with more than 10 and less than 20 seats	--
License plate color	Blue	Yellow	Black	Others
Document type	Insurance policy	Insurance cancellation	Information correction	--
Insurance type	Mandatory traffic liability insurance	Motor vehicle comprehensive insurance clauses	Shenxing auto insurance motor auto insurance	--

4 CASE: CUSTOMER SEGMENTATION IN AN AUTO INSURANCE COMPANY

In this section we demonstrate the practical value of our approach through a real case.

4.1 Data Collection and Preparation

Data for the case is about vehicle insurance sales from an auto insurance company, including comprehensive attributes of all vehicle insurances in the company since 2014. There are a total of 25738 objects, and each object is described with 46 attributes. Since some attributes do not have the value for clustering analysis, it is necessary to identify useful attributes. In the case, a total of 34 unrelated attributes are eliminated, including 18 sensitive information attributes, 5 attributes with unique value, 4 code number attributes, 4 unrelated date attributes, 3 meaning repetition attributes. Twelve valuable attributes are chosen and shown in Table 1.

In these attributes, seven attributes belong to categorical attributes including Policy nature, Policy state, Vehicle type, Seating capacity, License plate color, Document type and Insurance type, and they need to be digitized before further analysis. In particular, due to Seating capacity being described by digits originally, this attribute does not need to be digitized. The comparison table of categorical attributes digitization is shown in Table 2, and the categorical attributes listed in this table are only six except Seating capacity.

Additionally, the categorical and numerical missing values are imputed respectively by the mode values and

mean values of corresponding attributes. Specially, the imputed dataset is only used for k-prototypes and SBAC.

4.2 Customer Segmentation

Since improved k-prototypes and SBAC require higher time complexity, 500 data objects are extracted randomly as one group for experiments in this case, and each algorithm undergoes 20 groups.

In this case, we exploit Silhouette index (S) [21] to evaluate the effectiveness of clustering results. It is given as follows:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{9}$$

where $a(i)$ is the dissimilarity between data object i and its own cluster, and $b(i)$ is the dissimilarity between data object i and its neighbouring cluster. Obviously, the closer $s(i)$ is to one, the better the data object i is clustered. The average $s(i)$ over all data in dataset reflects the performance of clustering result.

(1) Customer segmentation using k-prototypes

In this case, k is given with nine values, respectively 2, 3, 4, 5, 6, 7, 8, 9 and 10, for each group, k-prototypes undergo a total of 180 runs. Silhouette index is utilized to select optimal clustering result in each group experiment for analysis.

Fig. 2 shows that the optimal clustering results of k-prototypes are scattered. Different initial prototypes lead to different optimal clustering results. Next, we will analyse several representative clustering results, which are selected from all optimal clustering results due to the limit of space. We select k-prototypes clustering results when the number of clusters (k) is equal to 2 and 7. These clustering results are respectively given in Tab. 3 and Tab. 4, in which each cluster is represented by representative data object.

Table 3 Clustering result of k-prototypes algorithm – k = 2

k=2	Cluster1	Cluster2
Policy nature	0	0
Policy state	0	0
Vehicle type	0	0
Seating capacity	5	5
License plate color	0	0
Document type	0	0
Insurance type	**1**	0
New car purchase price	**463195**	83289.39
Vehicle age	**0.92828**	1.943439
Accumulative paid-in amount	**7565.74**	1608.708
Written premium	**7565.74**	1608.708
Tonnage	0	0
object number	29	471

For auto insurance companies, the higher the order premium, the higher the profit for companies. Since 'Accumulative paid-in amount' and 'Written premium' can represent the order premium, these two attributes should be used to identify property of cluster.

In tables, the values that are worth being paid attention are in bold.

Table 4 Clustering result of k-prototypes algorithm – k = 7

k=7	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
Policy nature	0	0	0	0	0	0	0
Policy state	0	0	0	0	0	0	0
Vehicle type	0	0	0	0	0	0	1
Seating capacity	5	5	5	5	5	5	8
License plate color	0	0	0	0	0	0	0
Document type	0	0	0	0	0	0	0
Insurance type	0	0	**1**	0	0	0	0
New car purchase price	40263.3	120698.9	**358824**	198788.4	745696.7	70857.64	**28184**
Vehicle age	1.99333	1.737303	**1.04842**	1.824364	1.785	1.443493	**3.79543**
Accumulative paid-in amount	1082.38	1801.168	**4854.02**	3243.607	3877.767	1622.538	**958.6**
Written premium	1082.38	1801.168	**4854.02**	3243.607	3877.767	1622.538	**958.6**
Tonnage	0	0	0	0	0	0	**0.01429**
Data object number	81	89	19	55	12	209	35

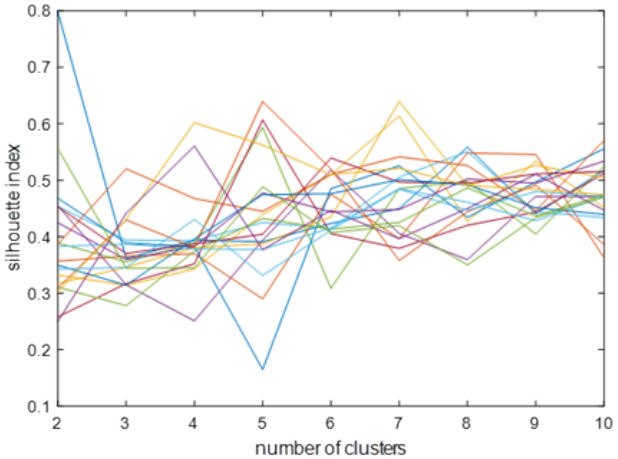

Figure 2 Silhouette index values of k-prototypes clustering results

(2) Customer segmentation using improved k-prototypes

For improved k-prototypes, k is also given with nine values respectively 2 to 10 for each group, and improved k-prototypes undergoes a total of 180 runs. Fig. 3 shows that the optimal clustering results are concentrated on k = 2 or 3 and very stable, that is related to the prototypes initialization method based on the number of nearest neighbours. Therefore, the randomness of selection is avoided, and the clustering result is more stable. Two optimal clustering results when k = 2 and 3 are respectively given in Tab. 5 and Tab. 6.

(3) Customer segmentation using SBAC

In this case, t is given with ten values, respectively 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09 and 0.1 multiple of the dissimilarity of root node $D(root)$, for each group. SBAC algorithm undergoes a total of 200 runs. Fig. 4 shows that the optimal clustering results of some groups

are mainly concentrated on $t = 0.02$, 0.03 and 0.04 multiple of $D(root)$. And the optimal clustering results of other groups appear when t takes 0.09 and 0.1 multiple of $D(root)$ that is shown in the bottom of the figure. However, since the Silhouette index values of the first situation are much larger than the second situation, we select the optimal clustering results when $t = 0.02$ and 0.03 multiple of $D(root)$ to exhibit respectively in Tab. 7 and Tab. 8.

Table 5 Clustering result of improved k-prototypes algorithm – $k = 2$

k=2	Cluster1	Cluster2
Policy nature	0	0
Policy state	0	0
Vehicle type	0	0
Seating capacity	5	5
License plate color	0	0
Document type	0	0
Insurance type	0	**1**
New car purchase price	67282.82	**158789.9**
Vehicle age	2.23488	**1.397799**
Accumulative paid-in amount	852.8802	**3570.466**
Written premium	852.8802	**3570.466**
Tonnage	0	**0.005789**
object number	291	209

Table 6 Clustering result of improved k-prototypes algorithm – $k = 3$

k=3	Cluster1	Cluster2	Cluster3
Policy nature	0	0	0
Policy state	0	0	0
Vehicle type	0	0	0
Seating capacity	5	5	5
License plate color	0	0	0
Document type	0	0	0
Insurance type	**1**	0	0
New car purchase price	**181747**	98958.19	65555
Vehicle age	**0.83919**	4.386443	0.67399
Accumulative paid-in amount	**4273.223**	1098.981	938.957
Written premium	**4273.223**	1098.981	938.957
Tonnage	0	0	0

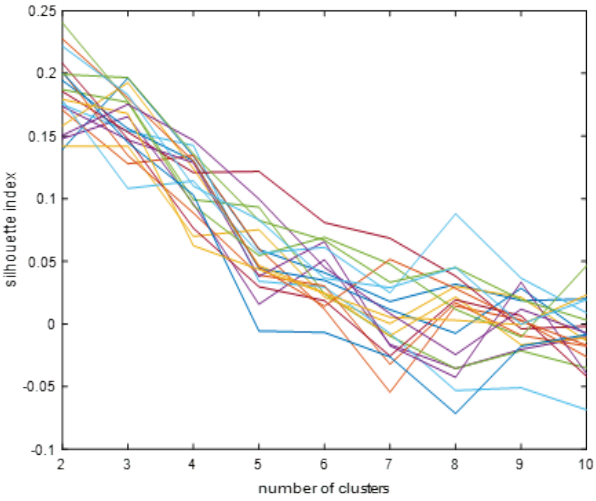

Figure 3 Silhouette index values of improved k-prototypes clustering results

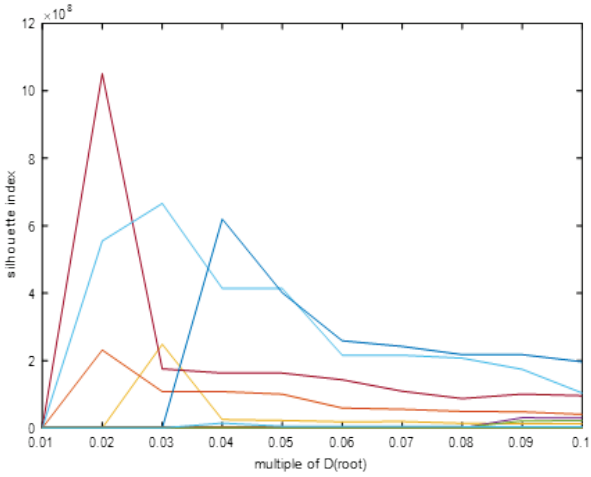

Figure 4 Silhouette index values of SBAC clustering results

Table 7 Clustering result of SBAC algorithm – $t = 0.02$ multiple of $D(root)$

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
Policy nature	0	1	1	0	0	0
Policy state	0	0	0	0	1	0
Vehicle type	0	1	0	0	0	0
Seating capacity	5	9	5	5	5	5
License plate color	0	0	0	0	0	0
Document type	0	0	0	0	1	0
Insurance type	0	0	2	**1**	2	2
New car purchase price	95544.08	**59800**	104732.1	**306405**	150313.3	195310
Vehicle age	1.908996	**6**	2.265455	**0**	2.64	4.8325
Accumulative paid-in amount	1900.8	**1055.97**	2989.338	**6942.06**	1083.542	5647.64
Written premium	1900.8	**1055.97**	2989.338	**6942.06**	1083.542	5647.64
Tonnage	0	0	0	0	0	0
Object number	468	2	11	2	6	4

Table 8 Clustering result of SBAC algorithm – $t = 0.03$ multiple of $D(root)$

	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
Policy nature	0	1	1	1	0	1	1
Policy state	0	1	0	0	0	0	0
Vehicle type	0	0	0	0	0	0	0
Seating capacity	5	5	5	5	5	5	5
License plate color	0	0	0	0	0	0	0
Document type	0	0	0	0	0	0	0
Insurance type	0	2	2	2	**1**	2	2
New car purchase price	102737.5	68273.33	97020	94320	**187020**	139660	175815
Vehicle age	1.814574	4.223333	2.5	4.5	**4**	2.71	2
Accumulative paid-in amount	1915.358	1839.56	2280.4	2563.33	**4832.86**	3496.06	3703.61
Written premium	1915.358	1839.56	2280.4	2563.33	**4832.86**	3496.06	3703.61
Tonnage	0.00252	0	0	0	0	0	0
Object number	481	3	2	2	2	2	2

4.3 Customer Characterization and Integration

In this subsection, we analyse the characteristic of each selected clustering result and extract rules by integrating clustering results of different algorithms. Furthermore, the effectiveness of each clustering result is discussed to verify the advantages of multiple algorithms analysis.

(1) The rules extracted from customer segmentation results

1) In the cluster corresponding to the highest accumulative paid-in amount, new car purchase price is always the highest.

Tab. 3 to Tab. 8 show that new car purchase price is always the highest when accumulative paid-in amount is the highest. Similarly, accumulative paid-in amount is also the highest when new car purchase price is the highest. It indicates that luxury car customers can produce more profits to insurance company, and they are important customers that should be followed by the insurance company.

2) In the cluster corresponding to the lowest accumulative paid-in amount, new car purchase price is the lowest.

Tab. 3 to Tab. 8 show that new car purchase price is always the lowest when accumulative paid-in amount is the lowest. Similarly, accumulative paid-in amount is also the lowest when new car purchase price is the lowest. It indicates that low-end car customers produce less profit to insurance company.

3) When the insurance type of one cluster is motor vehicle comprehensive insurance clauses, accumulative paid-in amount is always the highest.

Tab. 3 to Tab. 8 show that accumulative paid-in amount is always the highest when insurance type is motor vehicle comprehensive insurance clauses. It indicates that customers purchasing motor vehicle comprehensive insurance clauses always can produce more profits to the insurance company. Insurance company should focus on these kinds of customers, and actively recommend this insurance type to luxury car customers, thereby expanding this kind of customers group, and producing more income to the company.

4) When the vehicle type of one cluster is 6-10 seats, the accumulative paid-in amount is lower.

Tab. 4 and Tab. 7 show that accumulative paid-in amount is always the lowest or lower when vehicle type is 6-10 seats. 6-10 seat cars mostly belong to micro-bus, new car purchase price is also lower, and owners of this kind of cars will not purchase expensive insurance type, therefore they produce less profits to the insurance company.

5) In the cluster corresponding to the highest accumulative paid-in amount, the number of data objects is less.

Tab. 3 to Tab. 4 and Tab. 7 to Tab. 8 show that number of data objects is less when accumulative paid-in amount is higher. When accumulative paid-in amount is the highest, the new car purchase price is the highest and there are less customers that can afford to purchase luxury cars, so the number of data objects in the cluster with highest accumulative paid-in amount is less. However, Tab. 5 shows that the number of data objects corresponding to the cluster with high accumulative paid-in amount is not prominently less than the cluster with low accumulative paid-in amount, which is related to reduction of clustering result randomness in the improved k-prototypes algorithm. In addition, compared to the other two algorithms, the silhouette index value of improved k-prototypes algorithm clustering result is lower. Therefore, we think Tab. 3 to Tab. 4 and Tab. 7 to Tab. 8 are more convincing.

6) In the cluster corresponding to the lowest accumulative paid-in amount, the number of data objects is more.

Tab. 3 to Tab. 4 and Tab. 7 to Tab. 8 show that number of data objects is always the most or more when accumulative paid-in amount is the lowest. According to the second rule, customers who purchase low-end cars contribute low accumulative paid-in amount for the company. That is in line with reality, the number of customers purchasing low-end cars is prominently more than the number of customers purchasing luxury cars.

7) When vehicle ages are similar, the accumulative paid-in amount is high if the new car purchase price is high.

Tab. 4, Tab. 6, and Tab. 8 show that when vehicle ages are similar, accumulative paid-in amount is higher if the new car purchase price is higher. It indicates that insurance company should focus on luxury car customers when gaining new car customers.

(2) The effectiveness of different algorithm clustering results

Fig. 2 shows that the optimal clustering results of k-prototypes are more scattered. The optimal clustering results appear when the number of clusters k = 2, 4, 5, 6, 7, 8 or 10, which is related to the randomly selected prototypes. Fortunately, we can find different rules from the different results. Universal rules can be discovered according to the results with a smaller number of clusters, such as positive correlation between accumulative paid-in amount and new car purchase price. Detailed rules can be discovered according to the results with a greater number of clusters, for example, when the vehicle ages are similar, there is a positive correlation between new car purchase price and accumulative paid-in amount.

Fig. 3 shows that the optimal clustering results of improved k-prototypes are the most stable, which is distributed when the number of clusters k = 2 or 3. It is related to the prototypes initialization method. However, the scales of all clusters in improved k-prototypes algorithm clustering results are even, the scale of customers purchasing luxury cars and contributing more profits to the company are similar to that of customers purchasing low-end cars and contributing less profits. It is contradicted to the results of the other two algorithms about scales of all clusters. Since the silhouette index values of this algorithm clustering results are worse than those of k-prototypes and SBAC, we think the results of k-prototypes and SBAC are more accurate.

Fig. 4 shows that the optimal clustering results of SBAC are mainly distributed in the parts with smaller threshold t, the silhouette index values of the clustering results are higher compared with the other two algorithms. And the analysis of optimal clustering results of SBAC is consistent with that of other algorithms. Therefore, SBAC algorithm has better clustering effectiveness for auto insurance data in this case.

After analysing the clustering effectiveness, we can find that each customer segmentation result is affected by

the characteristics of the algorithm used. We cannot obtain convincing customer segmentation result if only one algorithm was used to do cluster analysis. Therefore, we utilize three clustering algorithms to segment customers, and the more convincing and accurate customer segmentation results can be obtained to support decision-making.

5 CONCLUSION

In this paper, we investigate how to mine more convincing and accurate customer related information of auto insurance companies through customer segmentation. Along this line, we propose an auto insurance business analytics approach for customer segmentation that exploits three mixed-type data clustering algorithms including k-prototypes, improved k-prototypes and SBAC to segment customers. In this way, the clustering results of these algorithms can complement and reinforce each other, and we can obtain as much information as possible to support customer related decision making of auto insurance companies.

The practical value of this work is confirmed in the fourth section, on one side, seven useful rules for the auto insurance company are extracted. Rules 1) 3) 5) and 7) show the characteristics of customers who produce more profits to the auto insurance company, and auto insurance company should actively expand the number of such customers, thereby contributing more profits to the company. Rules 2) 4) and 6) display the characteristics of a huge number of customers producing less profits to the company, and insurance company also should maintain them and convert them to another kind of customers that produce more profits. On the other side, the clustering effectiveness of different algorithms is also analysed in this paper to verify the validity of the approach. The case shows that the more convincing and accurate customer segmentation result can be obtained by utilizing multiple algorithms.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under Grant No. 71271027.

6 REFERENCES

[1] Matiş, C. & Ilieş, L. (2014). Customer relationship management in the insurance industry. *Procedia Economics and Finance, 2014*(15), 1138-1145. https://doi.org/10.1016/S2212-5671(14)00568-1

[2] Ka, Elan, L., Ka, Elan, V., & Novovi, Buri. (2014). A Data Mining Approach for Risk Assessment in Car Insurance: Evidence from Montenegro. *International Journal of Business Intelligence Research (IJBIR), 5*(3), 11-28. https://doi.org/10.4018/ijbir.2014070102

[3] Ke, N., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *Journal of Finance & Data Science, 2*(1), 58-75. https://doi.org/10.1016/j.jfds.2016.03.001

[4] Bhowmik, R. (2011). Detecting Auto Insurance Fraud by Data Mining Techniques. *Journal of Emerging Trends in Computing & Information Sciences, 2*(4), 156-162.

[5] David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics & Finance, 2015*(20), 147-156. https://doi.org/10.1016/S2212-5671(15)00059-3

[6] Kang, S. & Song, J. (2017). Feature selection for continuous aggregate response and its application to auto insurance data. *Expert Systems with Applications, 2017*(93), 104-117.

[7] Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: customer visit segmentation using market basket data. *Expert Systems with Applications, 2018*(100), 1-16. https://doi.org/10.1016/j.eswa.2018.01.029

[8] Ravasan, A. Z. & Mansouri, T. (2015). A Fuzzy ANP Based Weighted RFM Model for Customer Segmentation in Auto Insurance Sector. *International Journal of Information Systems in the Service Sector, 2*(7), 71-86. https://doi.org/10.4018/ijisss.2015040105

[9] Hanafizadeh, P. & Rastkhiz Paydar, N. (2013). A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry. *International Journal of Strategic Decision Sciences*, 1(4), 52-78. https://doi.org/10.4018/jsds.2013010104

[10] Du, M., Ding, S., & Xue, Y. (2017). A Novel Density Peaks Clustering Algorithm for Mixed Data. *Pattern Recognition Letters, 2017*(97), 46-53. https://doi.org/10.1016/j.patrec.2017.07.001

[11] Wangchamhan, T., Chiewchanwattana, S., & Sunat, K. (2017). Efficient algorithms based on the k-means and Chaotic League Championship Algorithm for numeric, categorical, and mixed-type data clustering. *Expert Systems with Applications an International Journal, 2017*(90), 146-167. https://doi.org/10.1016/j.eswa.2017.08.004

[12] Skabar, A. (2017). Clustering Mixed-Attribute Data using Random Walk. *Procedia Computer Science, 2017*(108), 988-997. https://doi.org/10.1016/j.procs.2017.05.083

[13] Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. In *1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 21-34.

[14] Li, C. & Biswas, G. (2002). Unsupervised Learning with Mixed Numeric and Nominal Data. *Knowledge & Data Engineering IEEE Transactions on, 14*(4), 673-690. https://doi.org/10.1109/TKDE.2002.1019208

[15] Goodall, D. W. (1966). A New Similarity Index Based on Probability. *Biometrics, 22*(4), 882-907. https://doi.org/10.2307/2528080

[16] Wu, S., Chen, H., & Feng, X. (2013). Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes. *International Journal of Database Theory & Application, 6*(5), 95-104. https://doi.org/10.14257/ijdta.2013.6.5.09

[17] Luo, H., Kong, F., & li, Y. (2006). Clustering Mixed Data Based on Evidence Accumulation. *Advanced Data Mining and Applications, 2006*(4093), 348-355.

[18] Chiu, T., Fang, D. P., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 263-268. https://doi.org/10.1145/502512.502549

[19] Cheung, Y. M. & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition, 46*(8), 2228-2238. https://doi.org/10.1016/j.patcog.2013.01.027

[20] Chen, H. L., Chuang, K. T., &Chen, M. S. (2008). On data labeling for clustering categorical data. *IEEE Transactions on Knowledge & Data Engineering*, 20(11), 1458-1472. https://doi.org/10.1109/TKDE.2008.81

[21] Rousseeuw, P. (1987).Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational & Applied Mathematics*, 20(20), 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Contact information:

Kai ZHUANG, PhD candidate
Donlinks School of Economics and Management
University of Science and Technology Beijing
30 Xueyuan Road, Haidian District, Beijing 100083, China
kyle_z80@yeah.net

Sen WU, PhD, Full Professor
(Corresponding author)
Donlinks School of Economics and Management
University of Science and Technology Beijing
30 Xueyuan Road, Haidian District, Beijing 100083, China
wusen@manage.ustb.edu.cn

Xiaonan GAO, PhD candidate
Donlinks School of Economics and Management
University of Science and Technology Beijing
30 Xueyuan Road, Haidian District, Beijing 100083, China
gaoxiaonan0001@163.com