Analysis of Preschool Linguistic Education Based on Orienting Problem Algorithm

Yan Zhao

Hanjiang Normal University, Shiyan, Hubei Province, 442000, China.

Abstract - This paper analyses mixed preschool linguistic education (CPBE) based on orienting convergence. The research steps are: i) we put orienting of fusion method system into mixed attribute data orienting problem, ii) we generalize the orient Ensemble Method, iii) apply the orient ensemble concepts to solve mixed attribute data orienting problem, iv) we establish the algorithm framework, v) put forward the objective function and algorithm, and finally vi) test the effectiveness of the algorithm on actual data.

Keywords - Orient problem-based; algorithm framework; preschool linguistic education

I. INTRODUCTION

Since China's preschool linguistic education sector join in the world trade organization (WTO), in the face of fierce market competition environment and the influence of foreign advanced technology and ideas, which bring about change and innovation. Under the current changing economy, preschool linguistic education most focuses on all kinds of information data collection and analysis, through the screening of optimal planning for all kinds of customer information, that provide differentiated way of financial products and services [1].

After joining WTO, the core competitiveness of preschool linguistic education has been converted into competition for high quality clients, preschool linguistic education industry’s demand for data center and customer analysis has been formed. For example: Industrial preschool linguistic education customer data’s screening is based on the customers basic services and credit risk, classified according to the customer repayment amount; Societe Generale, through customer profitability data analysis to classify, analyze the target customer base, marketing team takes corresponding strategy to expand customer base as per the analysis result; Other peers also analyze their own customers according to all kinds of data. Ping An subordinates will be officially joined the central preschool linguistic market credit reporting system. The core of credit reporting system is data collection, the greater the data samples, the stronger the regularity and the more value it owns. Ping An group has 74 million customers, accumulating of those personal information, which plays a key role for Ping An innovative financial Internet [2-3].

Customer segmentation is the premise of market segmentation and target marketing. Accurate market segmentation and differentiation marketing strategy is the difficult problem that the enterprise marketing to be faced with. The classification and orienting of data mining methods can be applied to customer group.

II. DATA ORIENTING AND ORIENTING ENSEMBLE

A. Data Orienting

Data mining orient is a group of individuals according to similarity to divide into several categories, makes the distance between individuals belonging to the same category as small as possible, and the distance between different categories of individuals as large as possible. Orienting is often used to develop targeted marketing according to dividing customer's behavior characteristics into different groups of users [4-5].

It’s not always get available result that revert the class attribute data into numerical attribute data, due to the value domain of the class attribute is out of order. Z.Huang proposed k - modes algorithm and k - as algorithm promote the k-prototypes method, which orient the class attribute and mixed attribute. However, by using of that way, which exist the problem of low level in exactness, stability and high randomness.

B. Orienting Ensemble

Orient ensemble is a rising new research field in the past two years. The basic idea is that with several independent orienters respectively on the original data, and group those results, finally get the orienting results of the original data. The actual data collection has the problem of irregular shape, noise, large amount of data and distributed data. However, by using of multi orienters, that can distribute processing data, in the meanwhile, the noise and outliers have less effect on the results, owns good stability, also has a good performance on processing irregular data and noise. For large data sets, using the appropriate orienters, which own a very good scalability as well .At the same time, the orient ensemble method can deal with the problem that single orient difficult to get the orienting result data set. At present most of the research literature has focused on numerical attributes, only class attribute is discussed, while the most widespread of mixed attribute data set problem in actual world has no articles involved [6-7].
III. CPBE ALGORITHM FRAMEWORK

CPBE mixed attribute data set algorithm framework, including numerical attributes and class attribute. Taking each class attribute as a orienting output result, each class attribute values are orienting label, the part of class attribute be converted to the same number of class preschool linguistic education results. Regarding numerical attributes, adopting the k-means forms orienting, using multiple orienting on these numerical attribute. Finally, taking the orienting result of class attribute and numerical attribute proceeds orienting ensemble, then the final orienting results are obtained. Algorithm framework as shown in figure 1.

Algorithm description: \( X = \{X_1, X_2 \ldots X_n\} \) presents a data set, \( X_i = \{X_{i1}, X_{i2} \ldots X_{im}\}, i = 1, 2, \ldots, n \) is data items, each data item has \( m = r + p \) attributes, which \( r \) presents the number of class attributes, \( p \) is the number of numerical attributes. Orienting is shown in \( H \), which map data set into a data item (or called the column label). For \( X \) class attribute shown in \( A_i \), \( i = 1, \ldots, r \), with domain \( V_i \). \( H_i \) is corresponding orienting of the class attribute \( A_i \), it map \( V_i \) to a natural number, complete the \( A_i \), division of a data item, obtain results \( \lambda_i \). The definition is shown as below:

\[
\lambda_i = \{H_i(x_i, A_i) \mid x_i, A_i \in V_i, x_i \in X\}, \quad i = 1, \ldots, r.
\]

Numerical attributes can be expressed with \( A_j, j=r+1, \ldots, m \). \( H_j \) stands for corresponding orienting of numerical attributes \( A_j \), it completes mapping for all the data item in each numerical attributes to orienting \( \lambda_j \), \( 1, 2, 3 \ldots m \) altogether the number of m orienting, through fusion function \( \Gamma \) fuse to \( \lambda \), gets the final orienting [8].

For example, table 1 contain eight items, each data item includes two class attribute and two numerical attributes. For class attribute \( A_1 \), according to the value of each data item, can be obtained results \( \lambda_1 = ((1,3,7),(2,4,5,6,8)) \). Classify the original eight items into two kinds, namely put data item 1, 3, 7 in the same class, item 2, 4, 5, 6, 8 in another; For \( A_2 \) class attributes, can be obtained result \( \lambda_2 = ((1,2,5,6),(3,7),(4,8)) \), classify the original eight items into three classes; Numerical attributes \( A_3 \), the results can be obtained \( \lambda_3 = ((1,2,8),(3,4,5),(6,7)) \), classify the original eight items into three classes; Numerical attributes \( A_4 \), can be obtained as the results \( \lambda_4 = (1,2,3),(4,5,6,7,8), \) classify the original eight items into two classes. In this way, the four \( \lambda_{1,2,3,4} \) could be obtained get the final result by using orienting ensemble [9-10].

### TABLE 1. DATA ITEM TABLE INCLUDING TWO CLASS ATTRIBUTE AND TWO NUMERICAL ATTRIBUTES.

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Class attribute</th>
<th>Numerical attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Z</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>Z</td>
</tr>
</tbody>
</table>

IV. OBJECTIVE FUNCTION AND COMMON ALGORITHMS

A. Objective Function

The goal of X orienting is to find a partition, incorporate all the data items in the X into the K misaligned group (K stands for a natural number). For n data items, it owns many feasible dividing methods, and the goal is to find a best, its need to define the objective function of the orient partition. Recently, many current orienting algorithms adopt the objective function based on distance or density. Here, according to the theory of Strehl, using maximize shared information as a objective function, Setting two partition of \( k_a \) and \( k_b \) respectively. In \( k_a \) \( k_b \) owns the number of \( k_a \) and \( k_b \) orienting respectively. In \( k_a \) orient \( C_h \), defines the sample number for \( n_h \); In \( k_b \), orient \( C_l \) defines the sample number \( n_l \). \( m^2 \) is sample number both in \( k_a \) h class and \( k_b \) l class. The normalized mutual information are:
Define $\Phi$ as Vector set of $k_{r+p}$, $\Phi$ and a partition $k_j$’s average normalized mutual information (ANMI) are:

$$
\Phi_{\text{ANMI}}(\Lambda, \lambda) = \frac{1}{r + p} \left[ \delta \sum_{i=1}^{r} \Phi_{\text{NMI}}(\hat{\lambda}_i, \lambda_i) + \sum_{j=r+1}^{r+p} \Phi_{\text{NMI}}(\hat{\lambda}_i, \lambda_j) \right]
$$

(2)

where, $i = 1, \ldots, r$ is corresponding partition of class attribute; $j = r + 1, \ldots, r + p$ are corresponding division for numerical attributes. Coefficient D is used to adjust the class attributes’ influence in orient, the problem will be discussed in another theory. (1) is the objective function. Set $K_{opt}$ for all partition, make the type (1) up to maximum partition,

$$
\lambda_{\text{opt}} = \arg_{\hat{\lambda}} \max \left[ \delta \sum_{i=1}^{r} \Phi_{\text{NMI}}(\hat{\lambda}_i, \lambda_i) + \sum_{j=r+1}^{r+p} \Phi_{\text{NMI}}(\hat{\lambda}_i, \lambda_j) \right]
$$

(3)

where $k_{d,l}$ is the partition of all possible options.

**V. ANALYSIS OF PRESCHOOL LINGUISTIC MARKET CUSTOMER RELATIONSHIP**

Preschool linguistic market credit card approval data in customer relationship management (CRM) include six numerical attributes and nine class attributes, a total of 690 items. The 15 attributes are: customer name, customer number, id number, gender, age, domicile of origin, employment status, income, industry, vocation, education, nature of the unit, the customer registration, customer classification, accumulated points. Product table mainly includes product type, name, details, customer profitability and corporate earnings and other information. Customer product associated data tables, mainly associate customers with its purchasing agent financial products, mainly include the following attributes: month, customer number, customer name, type, product name, monthly balance, the average monthly balance. Remove item 24 missing data, ultimately adopting 666 data items. The comparison of two kinds of orienting algorithms accuracy under different number of orienting as shown in figure 2.orienting results number from 2 to 8, respectively compare algorithm accuracy for each of the fixed orienting number. The average accuracy of k-prototypes is 0.77 k, CPBE average algorithm accuracy is 0.84.

![Fig 2. Comparison of the Two Algorithms](image)

Credit card authorization data put emphasis on the comparison of two kinds of algorithm accuracy, in order to verify the orienting performance. In the same set of data, fixed orienting number is 2. Respectively Random running two algorithms eight times, the comparison of two kinds of algorithm average accuracy are shown in table 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Prototypes</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>0.81</td>
</tr>
<tr>
<td>Second</td>
<td>0.82</td>
</tr>
<tr>
<td>Third</td>
<td>0.67</td>
</tr>
<tr>
<td>Fourth</td>
<td>0.78</td>
</tr>
<tr>
<td>Fifth</td>
<td>0.82</td>
</tr>
<tr>
<td>Sixth</td>
<td>0.75</td>
</tr>
<tr>
<td>CPBE</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>0.83</td>
</tr>
<tr>
<td>Second</td>
<td>0.84</td>
</tr>
<tr>
<td>Third</td>
<td>0.82</td>
</tr>
<tr>
<td>Fourth</td>
<td>0.83</td>
</tr>
<tr>
<td>Fifth</td>
<td>0.84</td>
</tr>
<tr>
<td>Sixth</td>
<td>0.83</td>
</tr>
</tbody>
</table>

CPBE orienting algorithm accuracy is relatively high. In eight times trials, CPBE algorithm accuracy changed little, the difference between the highest and lowest is 0.04. While
k-prototypes algorithm change a lot, the highest and lowest difference is 0.15. The reason for this is that k-prototypes algorithm is great affected by initial prototype selection. Because of multiple orienting ensembles, the CPBE algorithm orienting performance is relatively high and stable [14-16].

Customer segmentation is the premise of market segmentation and target marketing. Accurate market segmentation and differentiation marketing strategy is the difficult problems that enterprise marketing to be faced with. The classification and orienting of data mining methods can be applied to customer group. This project adopts the orienting method for customer segmentation, which provides a complete solution for postal financial marketing customer segmentation, and to verify its possibility with the actual case.

Here we consider customers orienting as per to the selected orient attributes. orienting algorithm realization process is as follows:

1) the selection of K value (this paper values: 12);
2) selecting initial centroid, as a heart of orients;
3) read each record, calculate the distance between the second record to orient heart, and attribute its distance to the nearest to orient heart, and then update the orient heart. Then repeat calculation the last third record;
4) and then repeat step 2 until the orient heart is no longer change

Orienting Results: Analyze customer classification after orienting, merge the orient with similar properties, eventually merge into 7 Orients.

For Orient 1 customer, marketing staff could focus on insurance, wealth management product sales, this part belongs to the high-end customers, and age is about 54, who pay attention to their own reimbursement and have certain financial management consciousness, is the main purchasing power for fund finance product currently.

Orient 2, elder customer, compare to other elder customer, this part belongs to high-end old customers. Their own more property relatively, if the marketing personnel have intention to guide them to buy financial products, insurance by bonus share is the most appropriate.

Orient 3 belongs to the rationally high-end customers, general depositors, can lead to buy some of the finance and insurance, but the difficulty is relatively high.

Orient 4 belong to the main purchasing power of the network of insurance products, but from the data analysis, buying insurance has a certain randomness, namely the marketing personnel success due to luck, and from the view of age structure, relatively belong to young group, are mainly distributed in 40 to 44 years old.

Orient 5 is the part of old people beyond 50 years old, they intend to savings, but you can recommend the appropriate insurance.

Orient 6 belongs to people who own assets under 2 w, focused on savings. It could guide to do financial management, such as fixed investment in funds

Orients 7 is losing customers.

For above seven kinds of customer, we obtain statistics for the customer distribution, 88.3% customers are mainly distributed in orients of 3, 4, 5, 6. With fewer high-end (orient 1, 2), for daily marketing outlets, suggest focusing on orient 3,4,5,6 introduce specific marketing strategy. In the notice of high-end customers, put emphasis on targeted marketing for the major customer in network.

Conduct initial analysis on customer data, then proceeds data pretreatment on agent financial products, both take customer’s purchasing difference of product and customer profit contribution for branch lines into consideration, classify financial customers, providing customers with differential financial products and services. For further application of big data, match the UnionPay card data with the customer, know customer daily consumption behavior, and excavate potential big customer.

Matching the customer info to the network, distributing the data, the network marketing team takes targeted strategy to expand the customer on the basis of data analysis.

VI. SUMMARY AND CONCLUSIONS

In the real world, especially business, data owns a large amount of mixed attribute variations. This paper introduced orienting ensemble method to mixed preschool linguistic education problems, put forward the mixed preschool linguistic education algorithm framework, determined the objective function based on orienting ensemble, and finally verified the process for the test data and customer relationship management (CRM) data. The experiment proved that CPBE algorithms have more improved results than k-prototypes in terms of accuracy and stability. Further research directions include: i) include fuzzy theory to further improve the algorithm for data processing ability and accuracy; ii) parallel algorithm study to explore the advantages of distributed computing.

ACKNOWLEDGMENT

This research is supported by the Education science Fund Project of Hubei Education Department (2014B554), the University Key Project of Hanjiang Normal University (2014004).

REFERENCES

YAN ZHAO: ANALYSIS OF PRESCHOOL LINGUISTIC EDUCATION BASED ON ORIENTING PROBLEM …


