IRSTI 28.01.75

D. Nurmambetov[1], A. Bogdanchikov[2]
[1,2]Suleyman Demirel University, Almaty, Kazakhstan

SEGMENTATION OF OUTLETS USING MACHINE LEARNING CLUSTERING AND CLASSIFICATION ALGORITHMS

Abstract. Segmentation of retail outlets in terms of manufacturing companies' strategy applied in sales amount and trade activities for each of them is very important. Directed investments into outlets help companies to make more profit and decrease expenses. This study presents a method which can be used for outlets clustering using unsupervised and supervised machine learning algorithms comprising 2 steps – data partition using unsupervised Gaussian Mixture (GM) Model clustering algorithm based on outlets sales amount and further partition for one of them using Logistic Regression (LR) and Neural Networks (NN) classification algorithms, which predict whether outlets will achieve monthly sales plan. Previously, clustering was made without any special methods and clusters were formed using some agreed threshold values for outlets sales amount. The proposed algorithm was tested on real sales data and formed 3 clusters according to business needs. Sales plan achievement prediction gave up to 74% accuracy.

Keywords: clustering, Logistic regression, Neural Networks, GM Model, prediction.

***

Аннотация. Сегментация торговых точек с точки зрения стратегии компаний-производителей, применяемой в рамках торговой деятельности для каждой из них, очень важна. Направленные инвестиции в торговые точки помогают компаниям получать больше прибыли и снижать расходы. В этом исследовании представлен метод, который можно использовать для кластеризации торговых точек с использованием алгоритмов машинного обучения с учителем и без, включающих 2 этапа - разделение данных с использованием алгоритма кластеризации модели Смеси Гауссовских распределений (GM) на основе объема продаж торговых точек и дополнительное разделение для одного из кластеров с использованием логистической регрессии (LR) и алгоритма нейронной сети (NN), которые предсказывают, будут ли торговые точки достигать ежемесячного плана продаж. Ранее кластеризация производилась без

каких-либо специальных методов, и кластеры формировались с использованием некоторых согласованных пороговых значений для объема продаж торговых точек. Предложенный алгоритм был протестирован на реальных данных о продажах и сформированы 3 кластера в соответствии с потребностями бизнеса. Прогноз выполнения плана продаж дал точность до 74%.

Ключевые слова: кластеризация, логистическая регрессия, нейронные сети, модель GM, прогнозирование.

***

Аңдатпа. Бөлшек сауда нүктелерінің әрқайсысы үшін сауда-саттық қызметі шеңберінде пайдаланылатын өндірістік компаниялардың стратегиясы бойынша сегменттеуі өте маңызды. Бөлшек сауда орындарына бағытталған инвестициялар компаниялардың пайдасын арттыруға және шығындарды азайтуға көмектеседі. Бұл зерттеуде сатылардың өткізу көлеміне негізделген 2 кезеңді қоса алғанда, оқытушымен және оқытушысыз компьютерлік оқыту алгоритмдері пайдалануанады - гаусс тарату (GM) кластерлік алгоритмі арқылы кластерлеу үшін пайдаланып, ал логистикалық регрессия (LR) және нейрондық желілік алгоритм (NN) пайдаланатын кластерлердің бірінде, олар сауда нүктелерінің ай сайынғы сату жоспарына жететінін болжайды. Бұрын кластерлеу арнайы әдістерсіз жүзеге асырылды, ал кластерлер бөлшек сауда орындарын келісілген сату шектік мәндер көмегімен құрылды. Ұсынылған алгоритм нақты сату туралы деректер бойынша сыналды және бизнес талаптарына сәйкес қалыптасты. Сату жоспарын іске асыру болжамы 74% дәлдігін берді.

Түйін сөздер: кластерлеу, логистикалық регрессия, нейрондық желілер, GM моделі, болжау.

Introduction

Retail outlets, also called trade channel, are the most important part of manufacturing companies' business, since all produced goods are sold to it. And outlets can be divided to several segments, which allows to implement certain actions to each of them.

The data used in this study are sales data of a certain category, consisting of company's sales data and sales of the whole industry members. Using this data, outlets can be divided into such clusters as:

- high – outlets with high sales amount, where company should hold their position

- potential – outlets with lower sales amount, but with a big potential to grow. It is about low company's sales and high amount of the whole industry sales.
- low – outlets with low sales amount, where company should attend or invest less.

The main steps are:

1. To create outlets clusters based on certain parameters, using clustering algorithms

3. To create a model, which can predict monthly sales plan achievement, which allows to get additional segmentation.

Literature review

The first step was to review machine learning general types and principles to choose appropriate algorithms and methods to the problem. There are many different types of machine learning algorithms, which can be useful depending on purposes and used data.

Supervised machine learning algorithms are those algorithms, which require algorithms' training before it can predict or classify outputs. The input dataset can be divided into train and test datasets. Cross-validation is one of the approaches, which is used to divide input data into training and test sets with the aim of avoiding overfitting. If the output of implemented algorithm is from continuous set, that kind of objectives are called regression problems, and it is called classification if the output set is discrete.

Supervised algorithms are used, if there are "correct" answers for the input data. Then classification accuracy can be estimated comparing "correct" answers with algorithm's output as share of correctly predicted from total number of answers.

Unsupervised learning algorithms do not learn any features from training data, and it is mainly used for clustering or decreasing dimension of data. [1]

The Neural Networks are based on a biological concept of neurons. A neuron is a cell in a brain, which communicates other ones by exchanging signals. NN include input and output layers and can also have several hidden layers. It can discover complex dependencies based on input data and find out non-linear functions to predict target data.

Proposed method

The preparation phase began with a selection of clustering algorithms to segment the points according to sales amount criteria. Many clustering algorithms and methods were considered during this study [3]-[5], since there is a huge variety, including the most popular K-means algorithm [6], as well as the Gaussian mixture algorithm [7] (an algorithm based on probabilities for

each sample of relation to each cluster) and complex models with ensembles of clustering algorithms [8].

The proposed method is to make two-step segmentation, including:

1) Using selected unsupervised clustering algorithm to create 3 main clusters.

2) Predicting whether outlet will achieve monthly sales plan based on additional features and patterns learnt from data.

A similar approach was used in [9], where students prediction problem was solved by partition of students and using specific algorithms or parameters for each part based on students' previous performance and behavior.

Data used at step 1 are company's sales amount and estimated sales of total industry for each outlet. Following algorithms were tested with real data and results were compared within step 1.

K-means finds the centers of the cluster, so that the distances from each point in the cluster to the center are minimal at each iteration. It stops, when cluster centers have not changed at the iteration. Susceptible to the choice of initial centers and noises in the data [10]. Three initial centers according to the business needs of the segmentation were chosen manually – high sales amount outlet, outlet with the potential and low sales amount outlet.

GM Model is a self-learning algorithm, commonly used for data clustering. Within it a statistical model is built — a normal distribution, that describes the data as accurately as possible, and then assigns the belonging of points to clusters.

It is worth noting, that data has little amount of noises and such outlets were dropped from the dataset. They can be classified after the main segmentation using K-nearest neighbors (KNN) algorithm [1].

For step 2 also variation of company's sales for outlet, outlet's past plans' achievement for the whole year and last 3 months were used as additional features. Sales plan achievement for current month was used as ground truth labels.

Then data was scaled with Standard Scaler to increase the accuracy and divided to train, test and validation sets. Proportion was 80% – train_val, 20% - test, where train_val was divided as 70% for testing and 30% for cross validation as it is the most common way for split. According to the purpose, accuracy was reviewed, and which is more important - precision, recall and F1 (harmonic mean of precision and recall) scores to track how many of positive predicted outlets were actual positive (precision) and how many of actual positive were predicted correctly (recall), as the model can be used for real business purpose.

Then Logistic Regression model and Neural Network were used to predict plans achievement and scores were compared. Similar approach was described for students' performance prediction in [11]. Multilayer Perceptron (MLP) Neural Network topology was used, since it is usually used for static data and for data, which is not too large for using complex topologies [2].

Also, it is important, that Polynomial features were added in LR model, since the data has non-linear dependencies to increase the accuracy [12].

Using cross validation method, the best polynomial power, regularization parameters for LR [13] and number of hidden layers and neurons, activation functions for NN [14-16] were defined.

Results and discussion

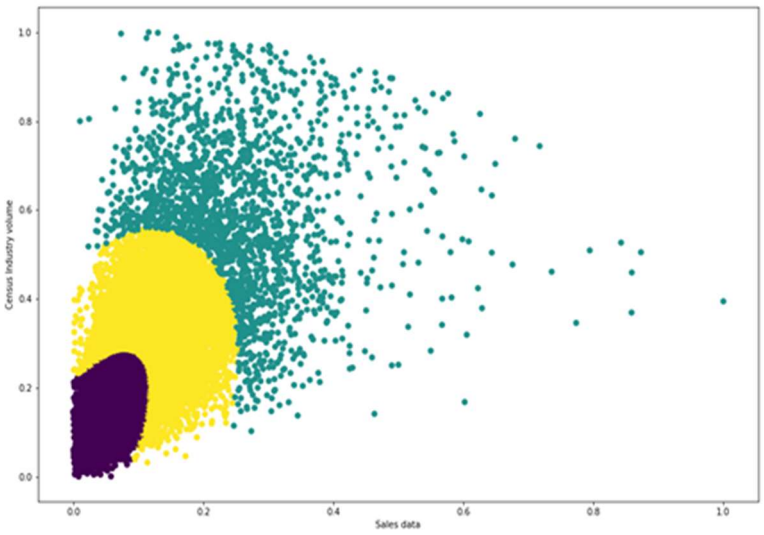

Fig. 1. GM Model clustering results

Since all clustering algorithms results can be estimated only visually, after comparing the results of clustered data by two algorithms, GM Model algorithm was chosen as more corresponding to business means of clusters, where every cluster presents company's strategy of investments in every outlet. Fig.1 shows the obtained clusters which were chosen because the 2nd cluster (yellow-colored) is more corresponding to outlets, which were mentioned above as "potential".

Also, the Silhouette Coefficient (SC) was used to estimate how well the both algorithms clustered the dataset. It is calculated using the mean distance

inside the cluster *a* and the mean nearest-cluster distance *b* for each outlet by the following:

SC= (b - a) / max (a, b)

The best value is 1 and the worst is -1. The results are shown in Table 1, K-means was implemented using 2 Python libraries – scipy and sklearn. All scores are similar, but K-means was not chosen despite better scores, because the obtained clusters had unsuitable form as mentioned above.

Table 1: SC for K-means and GM Model

K-means scipy	**K-means sklearn**	**GM Model**
0.457	0.465	0.416

As an additional estimation of clustering quality Fig. 2 shows the boxplot with distribution of outlets sales data by clusters formed with GM Model algorithm. It shows, that obtained clusters fit to the needs as described above – clusters differ with the amount of sales.

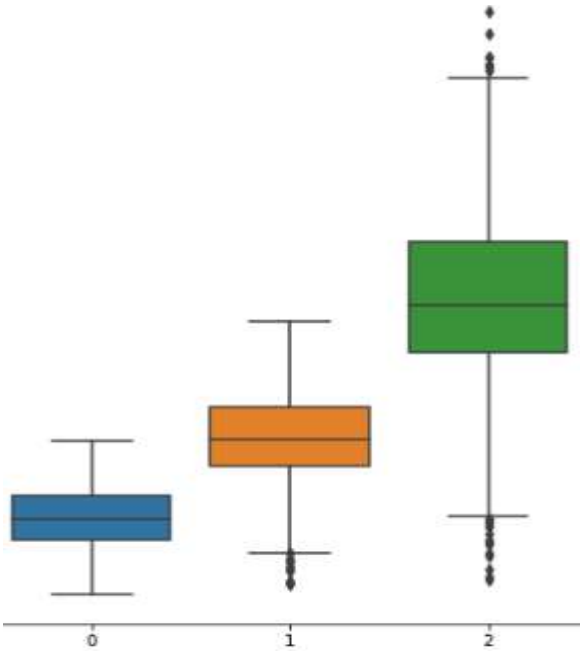

Fig. 2. Distribution of outlets sales data by clusters

During the next step LR and NN were used to predict sales plan achievement for low cluster as it can be clustered additionally.

As it was mentioned above, polynomial features were added to the data before fitting to LR model and to define best polynomial power, training accuracy for different powers was plotted. Fig. 3 shows, that 5 is the best choice as a global maximum of training accuracy. Feature engineering, including selection of most important features was also made to trade-off accuracy and computational speed.

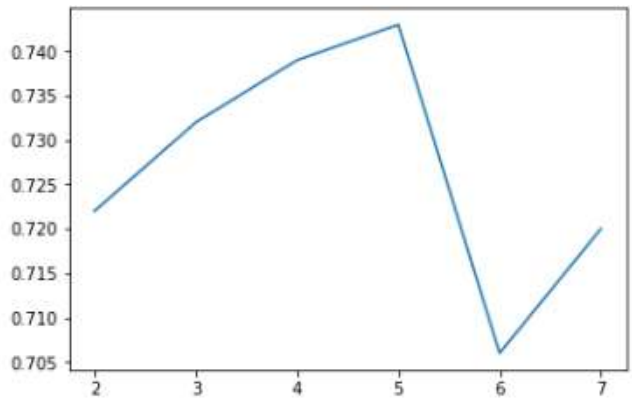

Fig. 3. Training accuracy of LR model depending on polynomial power

After using cross validation and defining best hyperparameters for the model, learning curve was plotted to estimate if the model underfits or overfits the data. Fig. 4 shows, that with increasing number of samples training and validation accuracy scores converge and are not too close. It means model is not underfitting or overfitting [17].

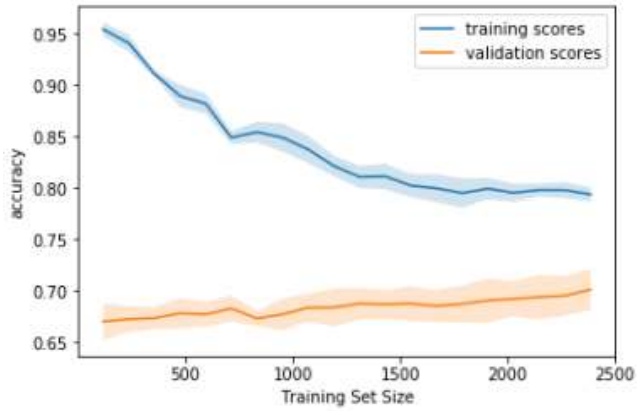

Fig. 4. Learning curve for LR model

For NN using Multilayer Perceptron the most important is to define the number of hidden layers and number of neurons in each. For this study Growing Method [2] was used, according to which model begins with small number of layers and neurons, then it increases and obtained results are compared to choose the best.

Most of problems and datasets does not need more than 2 hidden layers, and the number of hidden neurons should be less than twice the size of the input layer [18].

In Table 2 best scores after hyperparameters tuning for testing data using LR and NN are shown. NN gives slightly better accuracy and precision, but LR gives better recall. Precision and recall are very important in terms of business needs – trying to predict sales plan achievement with aim to focus efforts on such outlets.

Accuracy scores are like other works, using NN for predicting students' performance [2], [11] and show approximate potential of these models in terms of predicting plan achievement.

Table 2: Testing scores for LR and NN

Metric	**Logistic Regression**	**MLP Neural Network**
Accuracy	71%	74%
Precision	57%	62%
Recall	84%	71%
F1	68%	66%

Conclusion

In this study, outlets clustering method was presented, which can be used by manufacturing companies in terms of trade channel approach. Clustering results satisfy business needs and allow to use certain investments strategy for each outlet. Predicting sales plan achievement gave enough accuracy results, like other works solving such problem. But since predictions will lead to business decisions corresponding to it – it is very important to get higher precision and recall.

Using these models if they predict outlets as achieving plan, only around 60% will indeed achieve it, and only 84% of plan achieved outlets will be predicted correct. Also, considering that only 30% of the dataset have positive labels, such result isn't so valuable. Benchmark for this problem is >90% accuracy and F1.

Main advantage of predicting model is that it is not overfitting or underfitting. Its accuracy can be increased with creating ensemble clustering model or using additional features in the dataset. Proposed clustering method is already used and its second step with plan achievement prediction will be improved.

References

1 Dey, A. Machine Learning Algorithms: A Review. *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 7 (2016): pp. 1174-1179.

2 Oladokun, V., Adebanjo, A., Charles-Owaba, O. Predicting students' academic performance using artificial neural network: A case study of an engineering course. *The Pacific Journal of Science and Technology*, vol. 9, no. 1 (2008): pp. 72–79.

3 Berkhin, P. A Survey of Clustering Data Mining Techniques. In *Grouping Multidimensional Data,* by J. Kogan, C. Nicholas and M. Teboulle, ed. Berlin, Heidelberg: Springer, 2006. - P. 25–71.

4 Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31 (8), (2010): pp. 651–666.

5 Ayed, A. Ben, Halima, M. Ben and Alimi, A.M. Survey on clustering methods: Towards fuzzy clustering for big data. *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, *Tunis, Tunisia, Aug 11-14, 2014, IEEE:* pp. 331–336.

6 Forgy, E.W. Cluster analysis of multivariate data: efficiency vs interpretability of classifications, *Biometrics*, 21 (1965): pp. 768–769.

7 Kalti, K., Mahjoub, M.A. Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm. *The International Arab Journal of Information Technology*, 11 (1), (Jan 2014): pp. 11-18.

8 Ayed, A. Ben, Halima, M. Ben and Alimi, A.M. Adaptive Fuzzy Exponent Cluster Ensemble System Based Feature Selection and Spectral Clustering. *IEEE International Conference on Fuzzy Systems* (2017): pp. 1-6.

9 Li, Z., Shang, C. and Shen, Q. Fuzzy-clustering embedded regression for predicting student academic performance. *IEEE International Conference on Fuzzy Systems* (Jul 24, 2016): pp. 344-351.

10 Sukup, J. When K-Means Clustering Fails: Alternatives for Segmenting Noisy Data. Datascience.com. Feb 19, 2018. URL: https://www.datascience.com/blog/k-means-alternatives.

11 Arsad, P.M., Buniyamin, N., Jamalul-lail Ab Manan. Prediction of Engineering Students' Academic Performance Using Artificial Neural Network and Linear Regression: A Comparison. *IEEE 5th Conference on Engineering Education (ICEED)* (2013). DOI: 10.1109/ICEED.2013.6908300.
12 Agarwal, A. Polynomial Regression. Towardsdatascience.com. Oct 9, 2018. URL: https://towardsdatascience.com/polynomial-regression-bbe8b9d97491.
13 Kashnitskiy, Y. Open Machine Learning Course. Topic 4. Linear Classification and Regression. Medium.com. Feb 26, 2018. URL: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-4-linear-classification-and-regression-44a41b9b5220.
14 Cybenko, G. Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2 (4), (1989): pp. 303-314.
15 Hornik, K. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4(2) (1991): pp. 251–257.
16 Hinton, G. E., Osindero, S., Teh, Y. W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation,* 18 (7), (2006): pp. 1527–1554.
17 Luz, A. Why you should be plotting learning curves in your next machine learning project. Medium.com. Nov 26, 2017. URL: https://medium.com/@datalesdatales/why-you-should-be-plotting-learning-curves-in-your-next-machine-learning-project-221bae60c53.
18 Heaton, J. The Number of Hidden Layers. Heatonresearch.com. Jun 1, 2017. URL: https://www.heatonresearch.com/2017/06/01/hidden-layers.html.