P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i3.253

COMBINING BOOTSTRAPPING AND GENETIC ALGORITHM BASED ON FEATURE SELECTION FOR FRANCHISE LOCATION PROSPECT PREDICTION

Tati Mardiana

Program Studi Teknologi Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
tati.ttm@bsi.ac.id

Abstrak

Pemilihan lokasi sangat penting dalam industri waralaba makanan cepat saji. Model pemilihan lokasi menyeluruh yang dipasangkan dengan teknik analisis yang tepat dapat meningkatkan kinerja keputusan penempatan, menarik lebih banyak pelanggan, dan meningkatkan pangsa pasar dan profitabilitas. Kumpulan data lokasi waralaba memiliki sifat kelas yang tidak seimbang. Performa prediksi prospek lokasi waralaba menurun sebagai akibat dari karakteristik noise set data. Kami mengembangkan metode untuk meningkatkan kinerja prediksi prospek lokasi waralaba dalam penelitian ini. Ini menggabungkan Bootstrapping untuk mengatasi masalah distribusi kelas yang tidak seimbang dan algoritma genetika (GA) untuk memilih fitur yang relevan dalam prediksi prospek lokasi waralaba. Kami bereksperimen dengan lima metode klasifikasi yang berbeda (Naive Bayes, C4.5, Random Forest, ID3, Gradient Boosted Trees). Hasilnya menunjukkan bahwa hampir semua pengklasifikasi yang menggunakan Bootstrapping dan GA mengungguli teknik aslinya. Kami menggunakan Confusion Matrix dan Root Mean Squared Error (RMSE) untuk mengukur kinerja metode yang disarankan. Hasil pengujian menunjukkan bahwa metode kombinasi Bootstrapping dan GA sangat meningkatkan kinerja klasifikasi prospek lokasi waralaba.

Kata kunci: bootstrapping, algoritma genetika, lokasi waralaba, pemilihan fitur, prediksi.

Abstract

Location selection is crucial in the franchise fast-food industry. A thorough location selection model paired with a proper analytical technique can considerably improve the performance of placement decisions, attract more customers, and boost market share and profitability. Franchise location data sets have an imbalanced class nature. The franchise location prospect prediction performance decreased as a result of the dataset's noisy characteristics. We developed a hybrid approach to improve franchise location prospect prediction performance in this study. It combines Bootstrapping to address class imbalance problems and a Genetic Algorithm (GA) to select relevant features in the franchise location prospect prediction. We experimented with five different classification methods (Naive Bayes, C4.5, Random Forest, ID3, Gradient Boosted Trees). The results show that almost all classifiers that use Bootstrapping and GA outperform the original technique. We employ the Confusion Matrix and Root Mean Squared Error (RMSE) to examine the proposed method's performance. The test results demonstrate that the proposed method considerably enhances the franchise location prospect's classification performance.

Keywords: bootstrapping, genetic algorithm, franchise location, feature selection, prediction.

INTRODUCTION

Nowadays, franchising is a popular way for entrepreneurs to develop a business, especially when entering an intensely competitive industry like fast food. Owning a business remains everyone's dream. However, developing a business is uneasy and requires a mighty will. Entrepreneurs must generate business ideas and business plans, determine locations, product suppliers, hire employees, and carry out marketing strategies through various media. Conversely, if they decide on the wrong steps, it can lead to business failure. Ordinarily, entrepreneurs prefer franchise businesses to overcome obstacles in developing a business. Ordinarily, entrepreneurs prefer franchise businesses to overcome obstacles in developing a business. According to the International Franchise Association (IFA), franchising (Rosado-Serrano, Paul, & Dikova, 2018) is a type of license in which the parent organization, as the franchisor, gives the franchisee an independent entity in the form of the right to operate in a certain way. The franchisor typically provides an operational and marketing strategy manual, training and quality control, and provides business advisory support to franchisees. This helps reduce the risk of business failure as franchisees have a track record of success (Nguyen, Day, Wang, & Dang, 2017).

One of the most crucial factors of a franchise's success is its location. A good franchise location will attract more consumers and give convenient service(Chen & Tsai, 2016), which will increase customer loyalty. In addition, it possesses the potential to shorten the time to finance fixed capital investments, boost market share and profitability. Nevertheless, the business performance will decline if the location chosen for the business is not suitable.

The franchisor offers assistance to new franchisees in terms of location selection, in certain cases, lease negotiations. This assistance might be extremely beneficial in determining an ideal location. Ordinarily, before signing the franchise contract, the franchisee reserves the right to approve the location selected by franchisees. Nonetheless, determining the appropriate location is challenging because each alternative location possesses its own set of qualities (P, Widya Sihwi, & Anggrainingsih, 2016). Therefore, franchisees require tools to help determine the appropriate location according to the specifications set by the franchisor.

In recent years, companies have collected and preserved all information about business transactions in a massive database. It typically requires the implementation of alternative methods or technologies commonly referred to as big data. Data mining is computer science is the process of discovering and analyzing sizable amounts of data using conventional or semi-automated methods. With data mining, companies can obtain valuable information to increase their competitive advantage (Rao & Ramesh, 2019).

Despite the widespread use of data mining in various domains, only a few researchers utilize it for location selection problems(Chen & Tsai, 2016). In the past, researchers solved the problem of location selection using the C4.5 algorithms (Khumaidi, 2011), Rough Set Theory (Chen & Tsai, 2016), and Naive Bayes (Nuhayati, Dedih, & Mulyana, 2017) (Diana, 2017). However, franchise location data sets possess an imbalanced class nature. Imbalances can lead to impractical models in franchise location prospect prediction because these models predict most of the cases as a prospect. The accuracy of the franchise location prospect prediction model also decreased significantly because the data set contains attribute noise. It is necessary to utilize a method for selecting features and dealing with class imbalances to improve accuracy and minimize computation time.

The problem with imbalanced data distribution occurs when one class refers to the concept of interest which the number of negative instances outnumbers the number of positive instances(Feng, Huang, & Ren, 2018). Bootstrapping in machine learning is one of the methods used to deal with class imbalance, and it entails balancing (Naufal, Satria, & Syukur, 2015) the relative class distribution of data sets. Bootstrapping is the process of iteratively re-sampling a data collection with replacement to increase accuracy and reduce processing time.

Feature selection represents a method of reducing the dimension of the feature space and noisy data. In terms of feature selection, most feature selection algorithms try to explore solutions that fall between sub-optimal and near-optimal (Wahono & Herman, 2014). These algorithms implement a local search rather than a global search throughout the process, making it difficult to achieve near-optimistic solutions. The Genetic Algorithm (GA) can search the whole search space for a solution and execute global searches, greatly boosting the capability to provide high-quality solutions in a fair amount of time (Younas, Kamrani, Bashir, & Schubert, 2018)

We propose a hybrid approach to a franchise location prospect prediction model. It combines Bootstrapping for addressing class imbalance with GA for feature selection. Moreover, We conducted comparative research employing five classifiers to franchise location data sets in the context of franchise location prospect prediction. The proposed method achieves greater classification accuracy, according to the results of the experiments.

The following sections organized this paper. Introduce the proposed method for franchise location prospect prediction in section 2. Explain the study's findings and compare and contrast the methods suggested in section 3. Finally, explain the conclusions and suggestions of the study.

RESEARCH METHODS

Type of Research

This type of research is experimental research. We design and test models that deal with class imbalances and high-dimensional data sets to enhance accuracy.

Time and Place of Research

We properly researched for three months, from March until May 2021. We observed how to evaluate franchise locations conducted by franchise consultants from Khumaidi (2011).

Research Target / Subject

The population in this study is the franchise location that possesses done evaluated by a franchise consultant. This study's sample is a saturated sample, which includes the entire population. We utilized a sample of 120 fast-food franchise locations.

Research Procedure

The type of experimental design we utilize is a pre-experimental design with a one-group pre-test post-test design. The pretest-posttest group design is a research activity that includes giving a pre-test before treating the sample and giving a post-test after treating the sample. The following comprises the stages in this research.

1. Data Collection

   This stage collects data from previous studies on select franchise locations, criteria for selecting potential franchise locations and franchise location prospect data(Khumaidi, 2011).
2. Data Pre-Processing

   This stage cleans up data sets to eliminate inconsistencies, incomplete data, or redundant information.
3. Experiment and Testing Model

   In preference, We used classification methods to run trials using franchise location data sets. Following that, we apply the suggested method to data sets for franchise location prospect prediction, as in the pre-test, utilizing a classification algorithm.
4. Research Evaluation

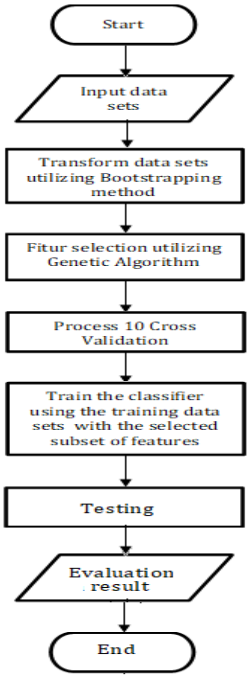
   At this stage, we present the results of the evaluation of the experiments that we possess done.

Figure 1 presents an activity diagram for the franchise location prospect prediction using a combination of bootstrapping and GA based on feature selection. The following is a process sequence of the methods proposed in this study (Agustian & Bisri, 2019).

1. Input the data sets contain training and testing data.
2. Transform the data sets utilizing Bootstrapping to measure the validity of data sets and give a balanced distribution by evaluating representative proportions of class examples.
3. After sampling the dataset, it selects the relevant features implementing GA.
4. For learning and testing data, we employ 10-fold cross-validation. We split the training data into ten portions with comparable characteristics and repeated the learning process ten times. We utilize stratified 10-fold cross-validation because it has become the industry standard in terms of practicality. According to certain studies, using shuffled enhances performance slightly.

   Train the classifier using the training data sets with the selected subset of features. Classify data sets with relevant features using two types of classification models, including traditional statistical classifiers (Naive Bayes (NB)) and Decision Tree (C4.5, Random Forest (RF), ID3, Gradient Boosted Trees (GBT)). Test the prediction model by looking at the accuracy and Root Mean Squared Error (RMSE).
5. Evaluation Models

   Validate if the proposed method produces a substantial divergence within the pre-test and post-test results. For each classifier pair, we performed a statistical t-Test without and with Bootstrapping and GA. The results show a more reliable proposed method for franchise location prospect prediction.


Start
Input data sets
Transform data sets utilizing Bootstrapping method
Fitur selection utilizing Genetic Algorithm
Process 10 Cross Validation
Train the classifier using the training data sets with the selected subset of features
Testing
Evaluation result
End

Source : (Mardiana, 2021)

Figure 1. Proposed Method

Data, Instruments, and Data Collection Methods

The following data support the implementation of this research.

1. Primary Data
   Primary data represents information obtained directly from sources such as interviews and confirmed observations at franchise locations.
2. Secondary Data
   Secondary data represents information gathered indirectly like paperwork, literature, books, journals, and other sources connected to the research subject.

The following are data collection methods in this study.

1. Observation
   Performing direct observations of the problems that occur in the selection of franchise locations.
2. Interview
   Interviewing franchisees about selecting a franchise location and confirming the findings of the data gathered from observations.
3. Literature Study
   Collecting literature, data and valuable information about franchise location selections from scientific journals, books, websites and magazines.

Data analysis method

The steps to analyze the data in this study are as follows:

1. Data collection
   Collecting franchise location data from previous research and compiling it in tabular form.
2. Selection and editing
   The data collected obtains raw data that contains incomplete, inconsistent or duplicate data. Therefore, we selected data and edited data whose measurement scales were inconsistent.
3. Data presentation
   Presenting experimental data in the form of diagrams or tables.
4. Conclusion
   The t-test results confirm that the proposed method significantly enhances the accuracy of the franchise location prospect prediction model.

RESULTS AND DISCUSSION

We employed franchise location data from previous research. We merely utilize data in the fast-food franchise category, as many as 120 records and 25 attributes. Prospect class attributes include three classes; very prospective with as many as 26 records, a prospective with as many as 70 records and a not prospective with as many as 24 records.

We started by experimenting with five different classification methods using franchise location data sets. It employs two various types of classification models, includes standard statistical (Naive Bayes (NB)) and Decision Tree (C4.5, Random Forest (RF), ID3, Gradient Boosted Trees (GBT)). Table 2 shows the results of the experiments. Testing the franchise location prediction model using NB resulted in an accuracy of 85 per cent with an RMSE value of 0.322. The C4.5 algorithms produce an accuracy of 70 per cent with an RMSE value of 0.434. Thereafter, RF produces an accuracy of 82.50 per cent with an RMSE value of 0.372. Meantime, ID3 produces an accuracy of 73.33 per cent with an RMSE value of 0.440. Finally, GBT produces an accuracy of 68.33 per cent with an RMSE value of 0.537.

The results explain NB and RF appear to be relatively well-performing methods for franchise location prospect prediction. The model used ID3 and C4.5 algorithms to produce relatively well-adequate results rather than GBT, which underperformed because of class imbalance. The model learning process time for each classifier is 0 seconds, except for the model that uses GBT for one second.

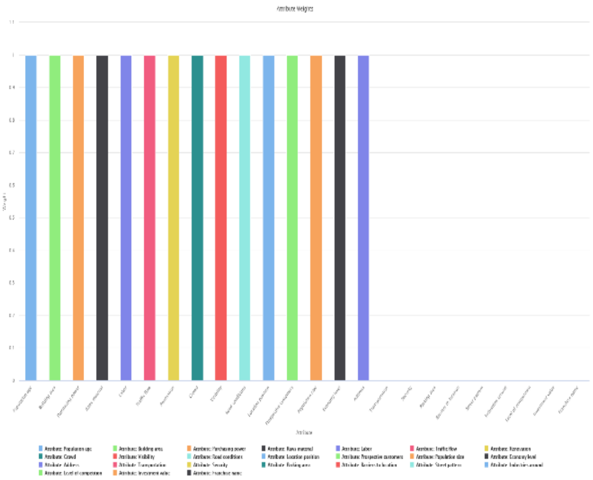
Table 1. Accuracy and RMSE of Classifiers (without Bootstrapping and GA)

Classifier		Accuracy	RMSE	Time Process
Standard statistical classifiers	Naïve Bayes	85.00%	0.322	0s
Decision Tree	C4.5	70.00%	0.434	0s
	Random Forest	82.50%	0.372	0s
	ID3	73.33%	0.440	0s
	Gradient Boosted Trees	68.33%	0.537	5s

Source : (Mardiana, 2021)

Figure 2 presents the attribute weighted of the franchise location data sets. The results of the selection process implementing GA obtained weight for each attribute of franchise location data sets. The attributes that weigh 1 are relevant attributes for the process of franchise location prospect prediction, including address, economic level, population size, location population, road conditions, visibility, crowd level, renovation, traffic flow, labour, raw materials, purchasing power, building area, and population age.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i3.253



Source : (Mardiana, 2021)

Figure 2. Attribute Weights of Franchise Location Data Sets

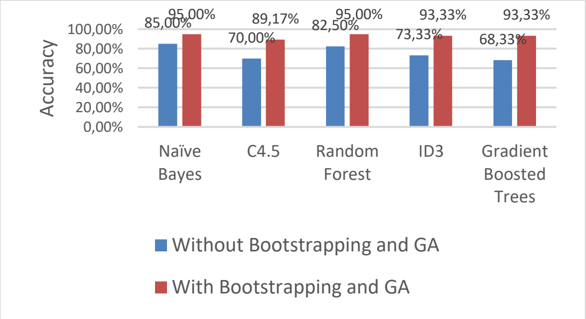
Table 3 shows the results of the experiment with Bootstrapping and GA. Testing the franchise location prediction model using NB resulted in an accuracy of 97.50 per cent with an RMSE value of 0.138. C4.5 produced an accuracy of 90.83 per cent with an RMSE value of 0.290. Then, RF produces an accuracy of 94.17 per cent with an RMSE value of 0.252. Meantime, ID3 produces an accuracy of 93.33 per cent with an RMSE value of 0.213. Finally, GBT produces an accuracy of 93.33 per cent with an RMSE value of 0.445.

Almost all classifiers that employ Bootstrapping and GA outperform the original technique, as demonstrated in Table 3 and Figure 1. It demonstrates that combining Bootstrapping and GA-based feature selection enhances classification performance significantly. The classification model that uses Bootstrapping and GA, on the other hand, takes longer to compute than the other.

Table 2. Accuracy and RMSE of Classifiers (with Bootstrapping and GA)

Classifier		**Accuracy**	**RMSE**	**Time Process (s)**
Standard statistical classifiers	Naïve Bayes	97.50%	0.138	6
Decision Tree	C4.5	90.83%	0.290	3
	Random Forest	94.s17 %	0.252	56
	ID3	93.33%	0.213	4
	Gradient Boosted Trees	93.33%	0.445	291

Source : (Mardiana, 2021)



Source : (Mardiana, 2021)

Figure 3. Comparisons of the accuracy of data sets classified by five different classifiers

Eventually, we compared the outcomes of the suggested approach (with Bootstrapping and GA) and as a method without Bootstrapping and GA to see if there is a substantial difference between the two ways. For every classifier pair without/with Bootstrapping and GA, we used the Statistical t-Test on each data set.

On the premise that the null hypothesis is true, the p-value in null hypothesis significance testing is the probability of generating test results that are at least as extreme as the actual results. The p-value is used to establish the smallest level of significance at which the null hypothesis is rejected. A lower p-value indicates that the alternative hypothesis is more strongly supported. To put it another way, the P-value indicates how likely the observed data is to emerge under the null hypothesis. The null hypothesis is rejected if the p-value is below the significant threshold (usually $\alpha < 0.05$), then the null hypothesis is rejected.

Table 4 shows the results of a two-tailed paired t-Test without/with Bootstrapping and GA. In the analysis, we obtained significant differences (P-value 0.05) across all five classifiers (Naive Bayes, C4.5, Random Forest, ID3, and Gradient Boosted Trees).

Table 3. The results of a two-tailed paired t-Test without/with Bootstrapping and GA

Classifier		**P-value of t-Test**	**Result**
Standard statistical classifiers	Naïve Bayes	0.000	Sig.(α<0.05)
Decision Tree	C4.5	0.000	Sig.(α<0.05)
	Random Forest	0.006	Sig.(α<0.05)
	ID3	0.000	Sig.(α<0.05)
	Gradient Boosted Trees	0.000	Sig.(α<0.05)

Source : (Mardiana, 2021)

CONCLUSIONS AND SUGGESTIONS

We have done research to deal with high-dimensional and imbalanced franchise location data sets. This study provides a new method for franchise location prospect prediction that combines Bootstrapping and GA. Almost all classifiers that use Bootstrapping and GA outperform the original method in the results. It demonstrates that combining Bootstrapping and GA-based feature selection enhances classification performance significantly. This application helps franchisees to determine an appropriate franchise fast-food location.

REFERENCES

Agustian, A. A., & Bisri, A. (2019). Data Mining Optimization Using Sample Bootstrapping and Particle Swarm Optimization in the Credit Approval Classification. *Indonesian Journal of Artificial Intelligence and Data Mining*, *2*(1), 18–27. https://doi.org/10.24014/ijaidm.v2i1.6299

Chen, L. F., & Tsai, C. T. (2016). Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, *53*, 197–206. https://doi.org/10.1016/j.tourman.2015.10.001

Diana. (2017). Sistem Pendukung Keputusan Menentukan Lokasi Usaha Waralaba Menggunakan Metode Bayes. *Jurnal Ilmiah Matriks*, *19*(3), 41–52.

Feng, W., Huang, W., & Ren, J. (2018). Class imbalance ensemble learning based on the margin theory. *Applied Sciences (Switzerland)*, *8*(5). https://doi.org/10.3390/app8050815

Hermawan, H., Fauzi, A., Cahyana, Y., & Handayani, H. H. (2020). Performa Optimal Penerapan Algoritma genetika Pada Penjadwalan Mata Kuliah. *Conference on Innovation and Application of Science and Technology (CIASTECH 2020)*, (02 Desember 2020), 683–690.

Khumaidi, A. (2011). Klasifikasi Data Prospektus Lokasi Waralaba Dengan Algoritma C.45. *Paradigma*, *Vol XIII N*(2 September 2011).

Mardiana, T. (2021). *Laporan Penelitian : Model Klasifikasi Prospek Lokasi Waralaba Makanan Cepat Saji*.

Naufal, A. R., Satria, R., & Syukur, A. (2015). Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan. *Journal of Intelligent Systems*, *1*(2), 98–108.

Nguyen, T. M. T., Day, J. Der, Wang, C. N., & Dang, H. S. (2017). Predicting of the performance of franchise industry using Grey models - Case study in United States. *Proceedings - 2017 International Conference on System Science and Engineering, ICSSE 2017*, (June 2020), 617–620. https://doi.org/10.1109/ICSSE.2017.8030948

Nuhayati, M. U., Dedih, D., & Mulyana, J. (2017). Sistem Pendukung Keputusan Untuk Menentukan Lokasi Usaha Kuliner Yang Strategis Menggunakan Metode Naive Bayes. *Jurnal Interkom: Jurnal Publikasi Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *12*(1), 4–12. https://doi.org/10.35969/interkom.v12i1.22

P, N. E., Widya Sihwi, S., & Anggrainingsih, R. (2016). Sistem Penunjang Keputusan Untuk Menentukan Lokasi Usaha Dengan Metode Simple Additive Weighting (SAW). *Jurnal Teknologi & Informasi ITSmart*, *3*(1), 41. https://doi.org/10.20961/its.v3i1.648

Rao, J. N., & Ramesh, M. (2019). A review on data mining & big data, machine learning techniques. *International Journal of Recent Technology and Engineering*, *7*(6), 914–916.

Rosado-Serrano, A., Paul, J., & Dikova, D. (2018). International franchising: A literature review and research agenda. *Journal of Business Research*, *85*(September 2017), 238–257. https://doi.org/10.1016/j.jbusres.2017.12.049

Wahono, R. S., & Herman, N. S. (2014). Genetic feature selection for software defect prediction. *Advanced Science Letters*, *20*(1), 239–244. https://doi.org/10.1166/asl.2014.5283

Younas, I., Kamrani, F., Bashir, M., & Schubert, J. (2018). Efficient genetic algorithms for optimal assignment of tasks to teams of agents. *Neurocomputing*, *314*, 409–428. https://doi.org/10.1016/j.neucom.2018.07.008