Gram-Schmidt Orthogonalization for Feature Ranking and Selection — A Case Study of Claim Prediction

Yuni Rosita Dewi, Hendri Murfi, and Yudi Satria

***Abstract*—Claim prediction is an important process in the insurance industry to prepare the right type of insurance policy for each potential policyholder. The frequency of claim predictions is highly increasing that head the problem of big data in terms of both the number of features and the number of policyholders. One of machine learning paradigms to handle the problem of the big data is dimensionality reduction by using a feature selection method. In this paper, we examine a new feature selection method for claim prediction using Gram-Schmidt Orthogonalization. In this method, the next features are iteratively selected based on the farthest distance to space spanned by the current features. Therefore, the advantage of the Gram-Schmidt Orthogonalization method is that it can provide a subset of the feature ranking without ordering all features. Our simulation shows that by using only about 26% of features, the predictor can reach comparable accuracy when it uses all features. It means that the Gram-Schmidt Orthogonalization-based feature selection method may need memory usage of about 26%, which is very significant in the context of the Big Data problem.**

***Index Terms*—Feature ranking, feature selection, Gram-Schmidt orthogonalization, big data, claim prediction.**

I. Introduction

The insurance industry is one of the industries that is growing very rapidly nowadays, almost everywhere. Both life and non-life insurance premiums increase by around 3% in Europe, North America, Asia-Pacific, and developing countries [1]. There are various types of non-life insurance, i.e., health, vehicles, home and property ownership, education, and much more. Insurance offers a more effective reduction in the level of anxiety because of risk and increases financial stability by mobilizing more efficient savings and capital allocations [2].

The frequency of claims tends to increase for some types of insurance. For example, the frequency of accident claims increased by 2.6% in the first quarter of 2014 until the first quarter of 2016 in Europe. This increasing frequency appears to be directly related to the increasing people who drive many miles away [3]. Therefore, claim prediction becomes one of the important processes in the insurance industry. By claim prediction, insurance companies can offer the right type and price of insurance policy for each potential policyholder.

Machine learning is a common method to solve the problem of claim prediction. For auto insurance, machine learning may predict if a candidate policyholder will initiate an auto insurance claim or not. In this case, the problem becomes a classification of supervised learning. There is usually a large amount of training data to construct the machine learning models. In big data terminology, the training data have a large volume in both the number of policyholders and features. There are some machine learning paradigms to handle the problem of the big data, especially volume context [4], one of which is dimensionality reduction by using a feature selection method.

Feature selection is a technique for determining the relevant features among the existing features in training data. A suitable feature selection method usually improves the accuracy of learning and make a better model. From the big data point of view, the feature selection may reduce the volume of the training data. It means that this approach provides some benefits, including the cost of building, storage, and processing models. There are several feature selection methods, including Simulated Annealing [5], Recursive Feature Elimination [6], and Naïve Bayes [7].

In this paper, we examine a new feature selection method using Gram-Schmidt Orthogonalization. The next relevant features are iteratively selected based on the farthest distance to space spanned by the current features. Therefore, the advantage of the Gram-Schmidt Orthogonalization method is that it can provide a subset of the feature ranking without ordering all features. Firstly, Arora *et al.* used Gram-Schmidt Orthogonalization to find the anchor words of separable nonnegative matrix factorization for topic detection in textual data [8]. Wang *et al.* applied Gram-Schmidt Orthogonalization to select words as features of textual data [9]. This work is quite like the work of Dewi and Murfi [10]; however, we use larger data from the problem of claim prediction. We examine the performance of the Gram-Schmidt Orthogonalization-based feature selection method based on the accuracy of Support Vector Machine Pegasos, which provides robust solutions [11] and suitable for larger data [12]. Our simulation shows that by using only about 26% of features, the classifier can reach comparable accuracy when it uses all features. It means that the Gram-Schmidt Orthogonalization-based feature selection method may need memory usage of about 26%, which is very significant in the context of the Big Data problem.

This paper is organized as follows: Section II describes the related research and the differences in this study from previous research. Section III explains the method used. Section IV shows the machine, the data, and the steps taken during the study. Finally, Section V presents the conclusion.

Manuscript received September 5; revised December 31, 2019.

Yuni Rosita Dewi is with Department of Mathematics, Universitas Indonesia, Indonesia (e-mail: yuni.rosita@sci.ui.ac.id).

doi: 10.18178/ijmlc.2020.10.1.898

II. Related Work

In 2004, Salcedo-Sanz researched feature selection methods with SVM to predict the bankruptcy of a firm or non-insurance company [5]. The feature selection method that he uses is Simulated Annealing (SA) and feature ranking using Walsh analysis. The data consists of 72 firms with 21 features. The result is that the feature selection algorithm combined with SVM is very good and useful when predicting the collapse of non-insurance companies in Spain.

Now we want to use case studies that use non-life insurance data, especially claim prediction problem. From the problem of the claim prediction, some researchers also have been used machine learning as a method of settlement. Weerasinghe *et al.* compared which machine learning method performs best in predicting the claim size of a policyholder. They compared three machine learning methods, i.e., neural networks, decision tree, and multinomial logistic regression. Their results indicated that neural networks were the best predictor [13]. Fauzan and Murfi also used XGBoost to predict claims from policyholders in the coming year by paying attention to the big data side [14].

Pozzolo compares various data mining models such as Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbors (KNN), Neural Network (NN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) for claim prediction [15]. This described research shows that basic methods, such as SVM can be used for problem claim prediction. For risk predictions in car insurance, Kašćelan uses nonparametric data mining techniques such as clustering, Support Vector Regression (SVR), and Kernel Logistic Regression (KLR) [16].

Insurance companies want the process of determining claims with a relatively fast time and accurate results. Feature selection in machine learning can be an alternative to reduce feature dimensions. In this paper, we use Gram-Schmidt Orthogonalization as a feature selection method. The advantage of the Gram-Schmidt Orthogonalization method is that it can determine the number of features without weighing all the features in the data.

Initially, Gram-Schmidt Orthogonalization was used by Arora *et al.* to look for anchors in Separable Nonnegative Matrix Factorization (SNMF) [8]. The data used is unstructured. Then Wang *et al.* used Gram-Schmidt Orthogonalization for feature selection in text data. In the study, the data used was also unstructured data [9]. Furthermore, Dewi and Murfi have used the Gram-Schmidt Orthogonalization method for small structured data [10]. It means in the previous study, Gram-Schmidt Orthogonalization is used to unstructured data, specifically to classifying text and grouping test. In this paper, we examine the Gram-Schmidt Orthogonalization for structured data like the work of Dewi and Murfi [10]. However, we use larger data from the problem of claim prediction.

III. Methods

This section explains the method used in this study. The method used as a feature selection method is the Gram-Schmidt Orthogonalization, and the method for calculating the accuracy of the sequence of features obtained is the Support Vector Machine (SVM) Pegasos.

A. Gram-Schmidt Orthogonalization

Gram Schmidt is the process of converting set vectors into orthogonal set vectors [17]. Arora is inspired by the work of gram-Schmidt, which uses the concept of perpendicular vectors to look for anchor words. This following algorithm by Arora to find the anchor word [8].

Algorithm 1. General Algorithm for Feature Selection

INPUT:
D: The structural text, containing M docs and W word.
K: The number of basis vectors (basis features).
T: The column dimension of random matrix R, and T << W

OUTPUT:
S: The *K* feature indices (indices of basic features)

Procedure : RP-GSO (*D, K, T*)

1. $[\bar{X},\tilde{X}]$ = generateWord_Doc_matrix(*D*) ;
2. $\bar{Q}$ = generateWord_Word_matrix($\bar{X},\tilde{X}$) ;
3. $\tilde{Q}$ = normr($\bar{Q}$) ;
4. R = generate_Random_Projection_Matrix(*T*);
5. $\overline{Q_{rp}} = \bar{Q} * R$;
6. S = gram_schmidt_orth($\overline{Q_{rp}}$, *K*):
   - 6.1 S = $\{d_i\}$ s.t. d_i is the farthest point from the origin;
   - 6.2 for i=1 to *K*-1
     Let d_j be the point in set *V* consisting of row vectors of $\overline{Q_{rp}}$ that has the largest distance to span(*S*);
     S ← $S \cup d_j$;
7. **return** *S*

The concept of Gram-Schmidt Orthogonalization that used in Algorithm 1 is modified to look for orthogonal projection from one point to a field. In this case, the point is a feature, and the plane is span(S). From the process of searching for an orthogonal vector, it is expected to obtain an ordered set based on important features to determine whether someone will submit a claim or vice versa.

The Gram-Schmidt Orthogonalization method used in the feature selection process in this study refers to Algorithm 1 with modifications according to the case studies studied. The following algorithm is used in this research. After that, to test the order of features obtained, the next step is a classification process described in subsection B.

Algorithm 2. Gram-Schmidt Orthogonalization as Feature Selection Method

Input:
D: Claim prediction data
K: The number of basis vectors (basis features).
Output:
S: the set of ordered *K* features
Prosedur:
V = *D*
S = gram_schmidt_orthogonal(*D, K*):

1. $S = \{d_i\}$ s.t. d_i is the farthest point from the origin;
2. V = V – d_i
3. For *i* =1 to *K*-1:
   Suppose d_j be the point in set *V* that has projected in span(*S*) that has the largest distance to span(*S*);
   $S \leftarrow S \cup d_j$
   $V \leftarrow V - d_j$

Return *S*

B. Support Vector Machine (SVM) Pegasos

SVM Pegasos is one method that applies online learning

and manipulation of the algorithm from Support Vector Machine as a basic method. All of the following explanation in this subsection referred to Shalev-Shwartz [12]. Suppose

$$S = \{(x_i, y_i)\}_{i=1}^{m}, \text{where } x_i \in \Re^n \text{ and } y_i \in \{+1, -1\} \quad (1)$$

with m is the number of the sample.

The problem is minimized:

$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{(x,y)\in S} l(w;(x,y)), \quad (2)$$

with loss function $l(w;(x,y)) = \max\{0, 1 - y\langle w,x\rangle\}$.

Run-time of algorithm SVM Pegasos does not depend directly on the size of the training data. The result of this algorithm is very suitable for learning from large datasets.

Step of SVM Pegasos:

1) Set w_1 into the zero vector. In the t-th iteration in an algorithm, select random samples from the training (x_{i_t}, y_{i_t}) data by taking the random uniform $i_t \in \{1,...,m\}$.
2) Replace the value in (2) with an approach based on training data samples (x_{i_t}, y_{i_t}) so that the form will be:

$$f(\mathbf{w};i_t) = \frac{\lambda}{2}||\mathbf{w}||^2 + l(\mathbf{w};(\mathbf{x}_{i_t}, y_{i_t})), \quad (3)$$

with λ is regularization parameter and loss function $l(\mathbf{w};(\mathbf{x},y)) = \max\{0, 1 - y\langle \mathbf{w},\mathbf{x}\rangle\}$.

Approach object values using sub-gradients with:

$$\nabla t = \lambda \mathbf{w}_t - \mathbf{1}\,[y_{i_t}\langle \mathbf{w}_t, \mathbf{x}_{i_t}\rangle < 1] y_{i_t}\mathbf{x}_{i_t}, \quad (4)$$

where $\mathbf{1}[\,y\langle \mathbf{w},\mathbf{x}\rangle < 1] y_{i_t}\mathbf{x}_{i_t}$ is an indicator function if the value is one, so the update weight will be:

3) Update $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla_t$

$$\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \eta_t 1[y_{i_t}\langle \mathbf{w}_t, \mathbf{x}_{i_t}\rangle < 1] y_{i_t}\mathbf{x}_{i_t}, \quad (5)$$

with learning rate $\eta_t = 1/(\lambda t)$.

The following algorithm of SVM Pegasos by Shalev-Shwartz [12].

Algorithm 3. Pegasos Algorithm

Input: S, λ, T
Initialize: Set $\mathbf{w}_1 = 0$
For $t = 1, 2, \ldots, T$
  Choose $i_t \in \{1,...,|S|\}$ uniformly at random.
  Set $\eta_t = \frac{1}{\lambda t}$
  If $y_{i_t}\langle \mathbf{w}_t, x_{i_t}\rangle < 1$, then:
    Set $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}_t + \eta_t y_{i_t}\mathbf{x}_{i_t}$
  Else (if $y_{i_t}\langle \mathbf{w}_t, \mathbf{x}_{i_t}\rangle \geq 1$):
    Set $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}_t$
  [Optional: $\mathbf{w}_{t+1} \leftarrow \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+1}\|}\right\}\mathbf{w}_{t+1}$]
Output $\mathbf{w}_{T+1}$

The advantage of the SVM Pegasos method is that the data retrieval process is done incrementally; that is, data processed in memory is taken partially. Then take the next part until all the data is used. When the next data collection, update weights such as (4) and (5) are carried out. Equation (3) in the case of a mini batch becomes

$$f(w;A_t) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{k}\sum_{i\in A_t} l(w;(x_{i_t}, y_{i_t})), \quad (6)$$

with k is the number of samples at each iteration, $1 \leq k \leq m$. m is the number of the sample, and $A_t \subset [m] = \{1,2,...m\}$, $|A_t| = k$. The following mini-batch algorithm on SVM Pegasos can be seen in Algorithm 4.

Algorithm 4. The Mini-Batch Algorithm

Input: S, λ, T, k
Initialize: Set $\mathbf{w}_1 = 0$
For $t = 1, 2, \ldots, T$
  Choose $A_t \subseteq [m]$, where $|A_t| = k$, uniformly at random.
  Set $A_t^+ = \{i \in A_t : y_i\langle \mathbf{w}_t, \mathbf{x}_i\rangle < 1\}$
  Set $\eta_t = \frac{1}{\lambda_t}$
  Set $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t\lambda)\mathbf{w}_t + \frac{\eta_t}{k}\sum_{i\in A_t^+} y_i\mathbf{x}_i$
  [Optional: $\mathbf{w}_{t+1} \leftarrow \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+1}\|}\right\}\mathbf{w}_{t+1}$]
Output $\mathbf{w}_{T+1}$

IV. Result and Discussion

This chapter explains the research process carried out, starting from the process of collecting data to the process of obtaining a sequence of features. Then an accuracy analysis process from each selected feature is performed to test the order of features obtained.

A. Data Sets

To build and evaluate the claim predictor, we use publicly available datasets from Porto Seguro through Kaggle (https://www.kaggle.com/c/porto-seguro-safe-driver-prediction). The training data is used to build a model as a predictor of probabilities a person will file a claim next year. Using the testing data, we estimate the accuracy of the model. Some elementary data information from these datasets are:

1) There are 595212 observations.
2) There are 57 features, and the target type is a binary label '0' for 'does not claim' and '1' for 'claim.'
3) Out of all values in the label, the comparison class between 'does not claim' and 'claim' is 26,44: 1. This comparison means the data has an imbalanced class.

```
Class 0: 573518
Class 1: 21694
Proportion: 26.44 : 1
```

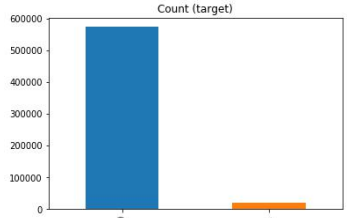

Fig. 1. The amount of data in each class.

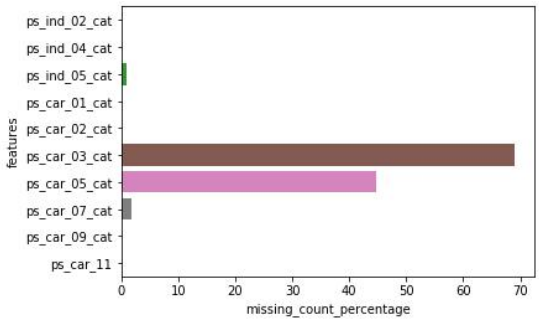

Fig. 2. Missing value in the data.

4) Feature titles follow the format 'ps_(group)_(n-th feature from the group)_(none/bin/cat)', where the groups are 'ind' features, 'car' features, 'reg' features, and 'calc' features. Features that are binary or categorical are given the tag _bin or _cat after the feature number. For example, ps_ind_03_cat is the third feature from the individual feature group, whose entries are of categorical values. The semi-black box nature of the features
5) There are massive amounts of NaN values in the dataset. The image below shows the percentage of missing value.

Data that has an imbalanced class and missing value must be completed first in the preprocessing step so that it can be processed to the next process.

B. Computer Software

The machines and computer software used to run the feature selection algorithm and classification of claim prediction in this study have the following specifications:

1) Operating system: Windows 10 64-bit
2) Processor: Intel(R) Core (TM) i5-8250U CPU @ 1.60GHz 1.80GHz
3) Installed memory (RAM): 8,00 GB
4) Software: Spyder 3.2.8 using Python 3.6.5

C. Preprocessing Data

The purpose of this stage is to make the data complete and ready for processing because not all the data was complete or ready to process. The data provided contains missing data so that the preprocessing data stage is needed, such as overcoming missing data using the most frequent (mode). The reason for using the imputer mode is that some of the missing value is categorical data.

Then a process is carried out to overcome the imbalance of data with random undersampling and random oversampling techniques. Random undersampling and random oversampling are one technique for handle imbalance class so that the class is a balance. The illustration of this technique is shown in Fig. 3.

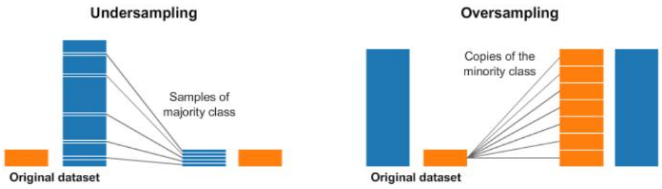

Fig. 3. The technique for handle imbalanced data.

```
Random under-sampling:
1.0    21694
0.0    21694
Name: target, dtype: int64
```

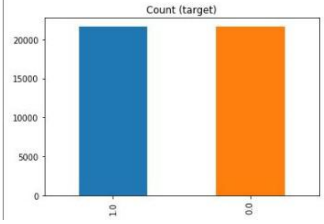

Fig. 4. Balanced data using undersampling technique.

After applied technique, it can be seen based on Fig. 1 that class '0' or the class 'does not claim' is a major class and class '1' or class 'claim' is a minor class. As a result, the data becomes a balance between the major class and the minor class with this technique. The balanced data is showed in Fig. 4 and Fig. 5.

```
Random over-sampling:
1.0    573518
0.0    573518
Name: target, dtype: int64
```

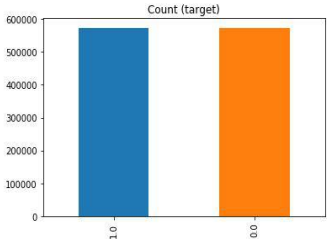

Fig. 5. Balanced data using oversampling technique.

D. Learning

The learning process aims to determine the parameters of the method in the training data provided. In SVM Pegasos, it takes one parameter, namely the parameters λ (lambda). The parameterλ is a parameter that is used to measure the trade-off from estimating the target training data. The value of λ is $\lambda = \{10^{-2},10^{-1},10^{0},10^{1},10^{2}\}$.

E. Evaluation

In this study, a model is said to be good in terms of the value of accuracy. If the value of accuracy in the evaluation process is high, the model can be used to replace the role of humans. For example, a set of data will be calcified into classes C_1 and C_2. The accuracy of the model testing is the ability to distinguish C_1 and C_2 classes correctly. To estimate the accuracy of a test, we must calculate the proportion of true positives, and true negative in all cases evaluated. Mathematically, it can be stated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \quad (7)$$

with:

TP (*True Positive)*: Numbers of true data that right classified as class C_1.

FP (*False Positive)*: Numbers of true data that wrongly classified as class C_1.

TN (*True Negative)*: Numbers of false data that right classified as class C_2.

FN (*False Negative)*: Numbers of false data that wrongly classified as class C_2.

F. Feature Selection

The result of feature selection using Gram-Schmidt Orthogonalization is the order or ranking of features. This sequence is expected to be a feature sequence based on the importance of the feature to determine whether a policyholder will make a claim or not. The order of the resulting features will be tested to be consistent. It was tested using data obtained with undersampling and oversampling techniques. The rank of 57 features in claim prediction data has been shown in the figure below.

```
feature_indices_list :

[31, 9, 26, 14, 46, 50, 21, 2, 47, 24, 44, 0, 49, 42, 4, 43, 45, 48, 41,
40, 29, 32, 36, 1, 53, 52, 3, 55, 15, 54, 20, 25, 5, 6, 22, 28, 56, 51,
23, 19, 7, 16, 39, 37, 38, 27, 18, 17, 13, 34, 30, 35, 8, 11, 33, 10, 12]
```

Fig. 6. The rank of feature from gram-schmidt orthogonalization using undersampling data.

We can see in Fig. 6 and Fig. 7 that the different order is in the 43rd and 44th feature, while the other is the same. It means that the order or the rank of features obtained is relatively consistent.

```
feature_indices_list :

[31, 9, 26, 14, 46, 50, 21, 2, 47, 24, 44, 0, 49, 42, 4, 43, 45, 48, 41,
40, 29, 32, 36, 1, 53, 52, 3, 55, 15, 54, 20, 25, 5, 6, 22, 28, 56, 51,
23, 19, 7, 16, 37, 39, 38, 27, 18, 17, 13, 34, 30, 35, 8, 11, 33, 10, 12]
```

Fig. 7. The rank of feature from gram-schmidt orthogonalization using oversampling data.

G. Accuracy of Feature Ranking and Selection Gram-Schmidt Orthogonalization Using SVM Pegasos

Based on the ordered feature resulted by Gram-Schmidt Orthogonalization, accuracy will be calculated using the number of features in the order of features obtained using SVM Pegasos. The simulation results show for undersampling data, the best lambda is $\lambda = 10^0$ and for the best oversampling lambda data is $\lambda = 10^{-1}$.

The following figure shows the accuracy of each feature using the ordered features. Fig. 8 shows the accuracy using the rank of feature in Fig. 6, and Fig. 9 shows the accuracy using the rank of feature in Fig. 7.

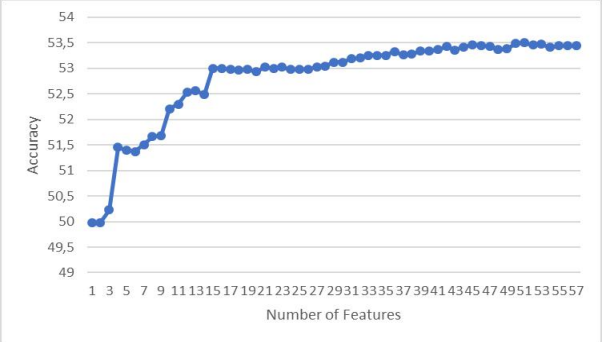

Fig. 8. Accuracy of claim prediction using undersampling data.

Based on Fig. 8, it can be seen the accuracy tends to rise, starting with using one feature up to 57 features even though there are decreases in some features. The highest accuracy occurs when the number of features is 51, which is equal to 53.50444%. This value is higher than 0.06316% of the total of 57 features. If noted, when the number of features 15, accuracy has reached 53.02535%. The difference between the number of features 15 and a total of 57 features is also less than one. So, we can use enough 15 features to determine whether a policyholder is likely to file a claim or not.

Based on Fig. 9, it can be seen the accuracy tends to rise, starting with using one feature up to 57 features even though there are decreases in some features. The highest accuracy occurs when the number of features is 55, which is equal to 56.83149%. This value is higher than 0.006721% of the total of 57 features. If noted, when the number of features 33, the accuracy has reached 55.9772%. The difference in accuracy between using 33 features and a total of 57 features is also less than one. So, we can use enough 33 features to determine whether a policyholder is likely to file a claim or not.

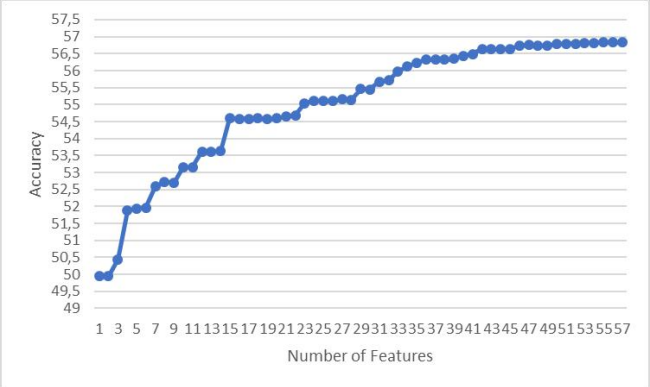

Fig. 9. Accuracy of claim prediction using oversampling data.

V. Conclusion

The feature sequence generated from the Gram-Schmidt Orthogonalization process for claim prediction problems is quite good. Based on the simulation, the resulting features are relatively consistent. The simulation also shows that by using only about 26% features, the resulting accuracy is comparable to using all features. It means that the Gram-Schmidt Orthogonalization-based feature selection method may save memory usage, which is very significant in the context of the big data problem.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The first author is tasked with cleaning data, doing simulation, and analyzing the result based on simulation. The second author is tasked with guiding, directing, and giving advice during the research.

Acknowledgment

This paper was supported by Universitas Indonesia under PIT 9 2019 grant. Any opinions, findings, conclusion, and recommendations are the authors' and do not necessarily reflect those of the sponsor.

References

[1] S. Crawford, L. Russignan, and N. Kumar, *Global Insurance Trends Analysis 2018*, U.K.: EY, 2018.

[2] M. Wieczorek-kosmala, "Non-life Insurance Markets in CEE Countries – Part I: Products' Structure," *J. Econ. Manag.*, vol. 25, no. 3, pp. 109–125, 2016.

[3] R. P. Hartwig and s. Weisbart, *More Accidents, Larger Claims Drive Costs Higher*, New York: Insurance Information Institute, 2016.

[4] A. L'heureux, K. Grolinger, and m. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776-7796, 2017.

[5] S. Salcedo-sanz, M. Deprado-cumplido, M. J. Segovia-vargas, F. Pérez-Cruz, and C. Bousoño-Calzón, "Feature selection methods involving support vector machines for prediction of insolvency in non-life insurance companies," *Intell. Syst. Accounting, Financ. Manag.*, vol. 12, no. 4, pp. 261–281, 2004.

[6] Z. Zeng, Y-W, Chen, C. Tao, and D. Van Alphen, "Feature selection using recursive feature elimination for handwritten digit recognition,"

in *Proc. Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009, pp. 1205-1208.
[7] M. Cinelli, Y. Sun, K. Best, J. M. Heather, S. Reich-Zeliger, E. Shifrut, N. Friedman, J. Shawe-Taylor, and B. Chain, "Feature selection using a one dimensional naïve bayes' classifier increases the accuracy of support vector machine classification of cdr3 repertoires," *Bioinformatics*, vol. 33, no. 7, pp. 951-955, 2017.
[8] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proc. the 30th International Conference on Machine Learning*, 2013, vol. 28.
[9] D. Wang, H. Zhang, R. Liu, X. Liu, and J. Wang, "Unsupervised feature selection through Gram-Schmidt orthogonalization-A word co-occurrence perspective," *Neurocomputing*, vol. 173, pp. 845–854, 2016.
[10] Y. R. Dewi and H. Murfi, "Feature selection using gram-schmidt orthogonalization for support vector regression – A case study of mortality rate prediction caused by pneumonia," *J. Phys. Conf. Ser.*, vol. 1192, 2019.
[11] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a technique for solvency analysis," *Ssrn*, no. 881, August 2008.
[12] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.
[13] K. P. M. L. P. Weerasinghe and M. C. Wijegunasekara, "A comparative study of data mining algorithms in the prediction of auto insurance claims," *European Int. Journal of Science and Technology,* vol. 5, no. 1, pp. 47–54, 2016.
[14] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," *Int. J. Advance Soft Compu. Appl*, vol. 10, no. 2, pp. 159-171, 2018.
[15] A. Dal Pozzolo, "Comparison of data mining techniques for insurance claim prediction," Thesis Universita Delgi Bologna, 2011.
[16] V. Kašćelan, L. Kašćelan, and M. N. Burić, "A nonparametric data mining approach for risk prediction in car insurance: A case study from the Montenegrin market," *Econ. Res. Istraz.* , vol. 29, no. 1, pp. 545–558, 2016.
[17] H. Anton and C. Rorres, *Elementary Linear Algebra*, 9th edition. New York: John Wiley & Sons, 2005.

Yuni Rosita Dewi received her bachelor of mathematics degree from Universitas Negeri Malang and Masters' degree in mathematics from Universitas Indonesia. Her research focused on the application of machine learning for big data.

Hendri Murfi received his bachelor of mathematics degree from Universitas Indonesia, master's in computer science from Universitas Indonesia, and Dr. Rer. Nat. from TU Berlin, Germany. Currently, he is serving as a lecturer and researcher in Data Science Group at the Department of Mathematics, Universitas Indonesia. He does research in machine learning with applications in topic modeling, sentiment analysis, recommender system, and insurance.

Yudi Satria received his bachelor of mathematics degree from Universitas Indonesia, master's in informatics from Institut Teknologi Bandung, and doctor from Universitas Indonesia. Currently, he is serving as a lecturer and researcher in Data Science Group at the Department of Mathematics, Universitas Indonesia. He does research in machine learning with applications in image processing and text analysis.