Analysis Using KNN and SVM

Ryutaro Takanami
Lancaster University, MSc Data Science

Abstract—**This paper aims to walk through the process of data science (load data-sets, pre-process the data, and apply various algorithms for analysis). To demonstrate clustering, hierarchical clustering and the K-mean algorithm are implemented for the pulsar data-set. After that, this paper shows two clustering algorithm, the K-Nearest Neighbor (KNN) and the Support Vector Machine (SVM) on two data-sets which has missing data and imbalanced data as well as showing pre-processing technique.**

I. Introduction

To explain the pipeline of this research, this paper is split in five sections. At first, this section introduces the outline of this research. The next section, Pre-processing, shows the several technique for dealing with data to apply the machine learning algorithm appropriately. In the Clustering section, the hierarchical clustering and the K-mean algorithm are described. The next Classification section demonstrates the results of both K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithm. After that, this paper concludes the results of analysis of three data-sets.

This research uses three data-sets, pulsar, mushroom and abalone. The pulsar data-set, which is used for clustering, describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey[1]. Pulsars are a rare type of Neutron star which produce radio emissions detectable on Earth. The data is 9 columns which has 17898 rows and all types of data are numerical. For classification, the mushroom and the abalone data-sets are used. The abalone data-set has 11 columns with 4172 rows which are numerical data without one categorical column. It is necessary to predict age of abalone with handling its imbalance. The last one, mushroom data-set, has 23 columns with 8125 rows and most types of data is categorical without only one column. All categories describe features of the mushroom and the aim is classifying whether a mushroom is edible or poisonous. This data-set has missing value in one column, then the method to deal with it is required.

In addition, this section describes the meaning of pre-processing, clustering and classification. Pre-processing is required to obtain the best outcome from machine learning algorithm through appropriate treatment for a data-set. For example, removing outliers which have an adverse effect for predicting, or deleting missing values because it leads to cause errors in the algorithms, if it is used straight[2]. Both clustering and classification are used for divide data into several categories but they are different at the point of the method of learning the data for predicting. Clustering is the unsupervised learning algorithm which means there are no answers for grouping data, but classification is the supervised learning which has labels to identify the answer of categories and use it for the learning process of algorithm[3]. The cross validation, which is the method to prevent the overfitting by split data into two data, the one is used to train and the other is used to test, is carried out after classification. The overfitting is the status which the predicted model is too fit to the data which is used to train, and cannot predict other data as well as the result which is demonstrated in the training[4].

II. Pre-processing

Pre-procession needs to modify the data into the best form for predicting. There are many methods like feature selection and/or exraction, and standardising and/or normalising. This process also provides insights into the structure and dependency of the data through exploring relationships or correlations[2].

A. pulsar data-set

For the three data-sets which were tackled in this research, different approaches were implemented because the appropriate pre- processing methods which is required to obtain the plausible results are different by the characteristics of data and algorithm[2]. About pulsar data-set, at first, standardising was carried out, then this research eliminated the outliers and normalized the data. Standardising is used to make sure the point of outliers (in this research, defined as the data in which variance is more/less than 3). Outliers should be eliminated because they are rare and too small/large value adversely affect to algorithms because outliers impact on the average value and scaling by normalising at the next pre-processing process. Normalising is necessary for scaling the data because the minimum, maximum, variance and mean of variable are different in each column and it leads to misreading in the learning process of algorithm. For instance, if the maximum value is much greater than the other columns, some algorithms like used in this report are affected the impact of the greater scale column, and the smaller scale columns are hard to be reflected to the prediction compare with greater one, because the algorithm uses distance of value for predicting[2].

$$x_{std} = \frac{x_i - \mu}{\sigma} \tag{1}$$

when x:data in column, x_{std}:standardized x

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2}$$

when x:data in column, x_{norm}:normalized x

B. abalone data-set

For the second data-set, abalone, oversampling and undersampling are implemented because this data is imbalanced data. Imbalanced data means that the data in the target variable which is predicted are almost same value, and the imbalance is problem because it makes hard to learn the minority data. In this case, predicting positive data in ”Class” column is required but it is only 32 of 4174 all data. Then, following methods aim to handle the amount of both the majority and the minority data.

Oversampling applies to the minority value to increase the rate of it by duplicating. This paper randomly sampled the minority value and duplicated one by one to 100. Undersampling means the reduction of the majority data, and implementing it with randomly sampling of it to the same amount of value as the over sampled minority value[2].

C. mushroom data-set

Handling missing values is the prime point of the mushroom data-set. As the method for it, inserting plausible value like the mean or median is used, but the data is categorical, then this report just deletes missing value alternatively and compares the result with the case the data inserted most frequent value and nothing for it (deal with missing value as a categorical value)[2].

Categorical variables cannot be used directly because the algorithm which are used in this report (KNN and SVM) require numerical values. There are two ways for converting, one-hot encoding and label encoding. One-hot encoding splits each category of the column to the different columns and express it as dummy variable which has only values of 0 and 1. However, as a demerit, it makes many dummy variables and needs to use many amount of the computing resource of computer like the memory. The other way, label encoding allocate the number to values in each category from 0. For example, when the column has three variables like ”apple”, ”banana”, ”melon”, the values are changed as ”0”, ”1”, ”2”. It does not expand the amount of columns like one-hot encoding, but algorithms like the KNN and SVM recognize the order of the numbers as distance[2]. Actually, there are no meaning in the order of numbers, when we use label encoding to the mushroom data-set. Then, this report uses one-hot encoding as converting method. One-hot encoding is also used in a column of abalone data-set for the same reason.

After one-hot encoding, the one category in each original feature are deleted to prevent from the dummy variable trap. For instance, in the mushroom data-set, there are “poisonous_e” and “poisonous_p” columns as dummy variables after one-hot encoding because the original categorical column “poisonous” has two kinds of the value, e and p. Then, the one of the dummy variables in a categorical column is deleted. The dummy variable trap is the problem that dummy variables which produced by one-hot encoding tends to be high correlated, multicollinear. It means the one variable can be predicted with other variables in the model. The reason of it is the one dummy variable can be calculated by the other dummy variables which produced from the same categorical variable. For example, the values in “poisonous_e” can be computed by change all values in “poisonous_e” (by changing 0 to 1 and 1 to 0). The predicted values by using high correlated data can be doubted the reliability of the outcome because they adversely affect the standard error which shows the dispersion of data. Therefore, the one of the dummy variables in a categorical column is deleted[5].

III. Clustering

Mainly, there are two methods for clustering, the hierarchical and non-hierarchical methods. Hierarchical method starts from recognizing each data as a cluster at first. Then, it merges the clusters from the closest data one by one until merging all clusters. There variety of method to calculate the distance between each cluster, and this report uses euclidean distance and ward’s method. The equation of these are below[3].

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \tag{3}$$

$$d(p,q)=E(p_i\vee q_i)-E(p_i)-E(q_i) \tag{4}$$

when $E(p_i)=\sum_{x\in p_i}(d(x,q_i))^2$

The other way, the non-hierarchical method, defining the cost function and split data to cluster as many as the number pre-decided with the way minimizing the cost function. Compare with hierarchical method, it is different that the number of clusters is decided. This paper uses K-mean method which minimises the distance between the each centoid, which is computed by each cluster, and each data of the cluster. In each step of computing the distance, K-mean allocates each data the label of each cluster which is the same label as the closest centroid. After that, the algorithm updates the centroid with each clustered data and allocates the label to each data again. If centroids does not move the point from the before step, the clustering is finished. K-mean can be define the number of the cluster with the parameter k, and this research define it as 2 to divide into pulsar data and non-pulsar data. This paper compared the accuracy rate of each clustering algorithm to examine the algorithm divided into two clusters correctly. The result of euclidean distance, ward’s method and K-mean method were 94.5%, 74.2% and 95.2% each.

The accuracy of euclidean distance and ward’s method are different by about 20%. To confirm the difference of the clustering, the dendrogram, which illustrates the structure of the hierarchical tree, is useful. As we can see from the dendrograms, when the data is split into two clusters, the amount of data in each cluster is different. The amount of data in the cluster colored green in the dendrogram of euclidean distance is smaller than that in the dendrogram of word’s method. This can be considered as the reason of the accurate difference, and there are 967 positive target values in the 16172 rows. Therefore, the cluster which classifies the pulsar data should be smaller than the other cluster.

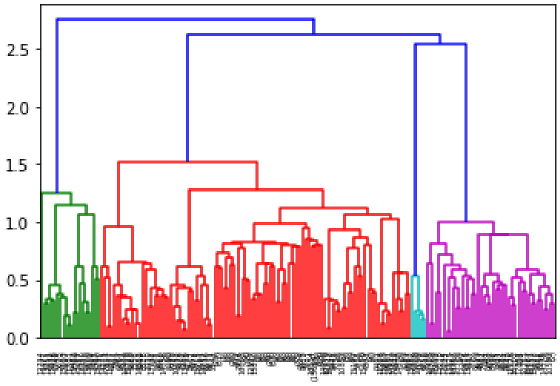

Fig. 1. the dendrogram of euclidean distance

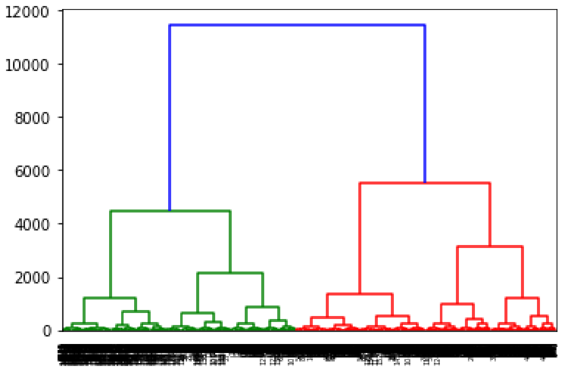

Fig. 2. the dendrogram of ward’s method

In addition, Figure 3 illustrates the results of the clustering by K-mean. Each dot in this plot is each data in pulsar data-set ,and the data which clustered as pulsar data is coloured red and the other one, non-pulsar data, is coloured blue. Centroids of each cluster are marked as x. Originally, the data has 8 dimensions, then dimensions were reduced by Principle Component Analysis (PCA). PCA reduces dimensions by transforming data to new features. From Figure 3, each centroid was far each other enough to divide data into 2 clusters and data were divided clearly.

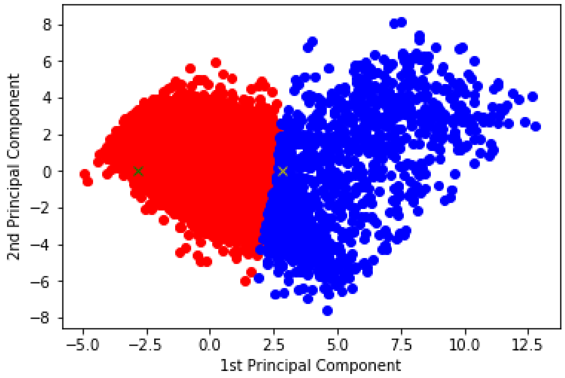

Fig. 3. The result of clustering by K-mean

IV. Classification

Classification is the supervised machine learning algorithm which classifies the data with labels. This paper used two algorithm, the KNN and SVM.

The KNN classifies new data by learning the trained data. When the KNN receives the new data, KNN compared it with the close trained data, and predicts its label as the majority of close data has. In order to calculate the distance between the new data and the trained data, euclidean distance is used and the number of trained data which is taken account as close data depends on the parameter k. That is why this algorithm is called K-Nearest Neighbor[3]. SVM creates the line which maximizes the distance from each class[3].

This report applied both algorithms to two datasets, abalone and mushroom, to examine the difference of the accuracy rate and the contribution of pre-processing like undersampling or the technipues for handling missing data to the prediction.

A. Abalone

In this sub-section, this report optimize the parameter k of KNN at first, and examine the change of accuracy rate by the percentage of negative class by both the KNN and the SVM.

Figure 4 represents the result of the change the parameter k. The accuracy rate decreased from the beginning to around 25 nearest neighbours, then dramatically increased to about 50 nearest neighbours. After that, the increase became slight and stayed the same from around 175 nearest neighbours. From this result, the label of neighbours tends to be same mostly and the data was classified well because accuracy rates were more than 90% in all case. However, the neighbour which is at between the first and the 25 from the predicted point tends to have the different label.

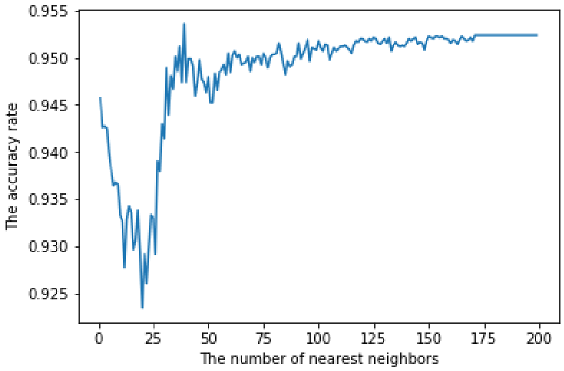

Fig. 4. The number of nearest neighbor and the accuracy rate on the abalone data-set (the percentage of negative values is 200)

The comparison of the accuracy rate with variety of percentage of the majority class (in this case, the label is negative) between the KNN and the SVM is Figure 5 and Figure 6. This report implemented KNN with defining 175 as the parameter k. The shape of the change were similar since the accuracy rate sharply dropped from around 90% when the amount of negative rate less than positive rate, and logarithmically increased from the bottom of the accuracy rate after that. The different points in each plot were the bottom of the accuracy rate and the way to recover it. The bottom of the accuracy rate by the KNN was less than by SVM but the accuracy rate increased sharply compared it with the result of the SVM. From these results, the number of undersampling should be more than the number of the minority data and the more the number of undersampling increase, the more the accuracy rate improve.

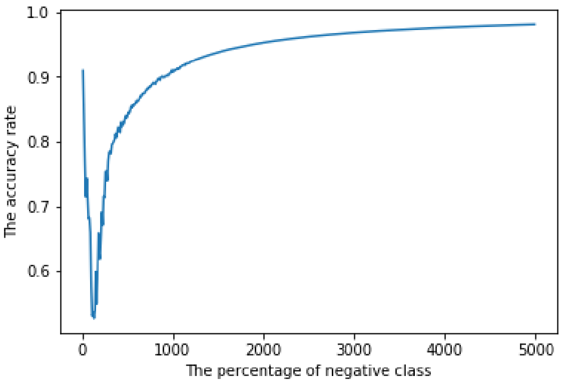

Fig. 5. The percentage of negative values and the accuracy rate on the abalone data-set (KNN)

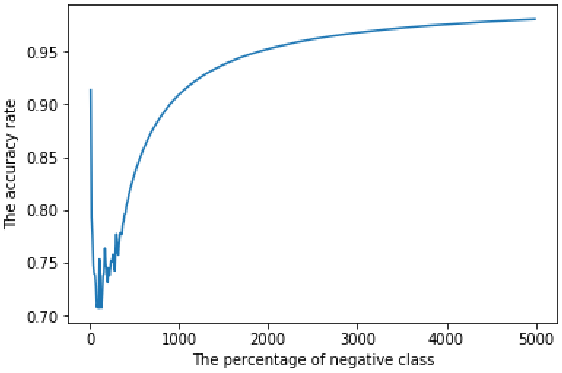

Fig. 6. The percentage of negative values and the accuracy rate on the abalone data-set (SVM)

B. *Mushroom*

About the mushroom data-set, this report optimize the parameter of KNN in each handled data which used different approach of pre-processing. After that, the accuracy rate by SVM in each handled data is examined, and compared with the outcomes by KNN. As well as it is explained in the pre-processing section, there are three handled data which are deleted missing values, inserted the most frequent value (in this case, the value is “b”) and handled missing values as another category.

The results of the optimization of KNN in each handled data are Figure7, Figure8 and Figure9. Shapes of the each result were similar and accuracy rates were decreased from 1.0 (100%) but the accuracy rate of the handled data which was inserted the most frequent value declined gently compared it with other two results of handled data because the accuracy rate was recovered to around 0.9985 when the number of nearest neighbours was 50. Moreover, the accuracy rate of the handled data with deleting was the lowest. The reason of these results can be analysed that the handled data by deleting lost the data which could predict the target variable accurately. Another method, handling missing values as new categorical value, contributed to improve the accuracy rate compare with the data which deleted missing values because it had more data which could be used to predict and expressed the case which produce missing value by inserting the new categorical value. However, the missing data seems the data in each categorical value in the same column (“b”, “c”, “u”, “e”, “z” or “r” in the column) and the amount of “b” is about two third of the data without missing value. Therefore, two third of missing values seem to be “b” and the handled data by inserting the most frequent value obtained the best accuracy rate.

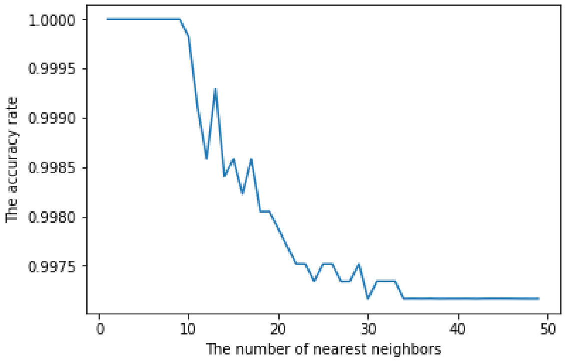

Fig. 7. The number of nearest neighbor and the accuracy rate on the mushroom data-set (handled missing values with deleting)

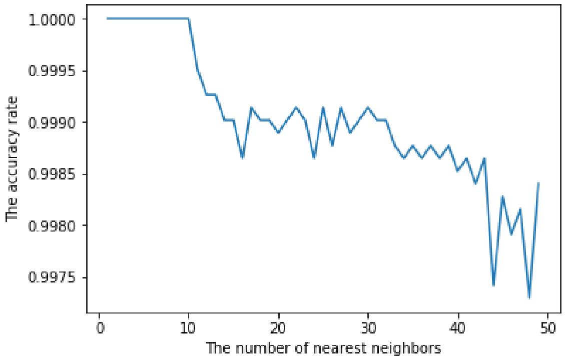

Fig. 8. The number of nearest neighbor and the accuracy rate on the mushroom data-set (handled missing values with inserting the most frequent value)

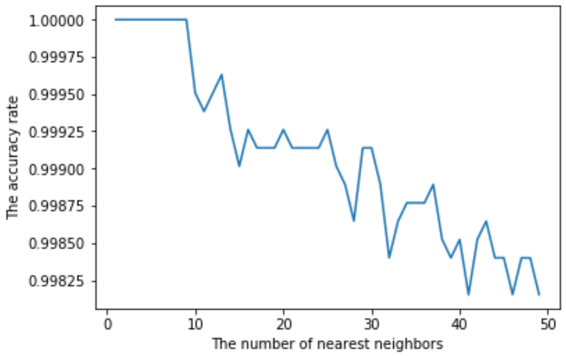

Fig. 9. The number of nearest neighbor and the accuracy rate on the mushroom data-set (handled missing values as a new categorical value)

Accuracy rates by SVM in each handled data were all 1.0. The accuracy rate is the same as the accuracy rate by optimized KNN (the number of nearest neighbours is from 1 to around 10). Both accuracy rates were high, but these were not over fitted results because the cross validation was implemented with 5 folds. Therefore, these results can be concluded that the number of features and data were enough to classify whether a mushroom is edible or poisonous.

V. Conclusion

In conclusion, this paper analysed three data-sets, pulsar, abalone and mushroom, and applied various techniques of pre-processing appropriately to each data-set, and algorithms of clustering and classification. In addition, the difference of results by using different pre-processing techniques and algorithms were examined. Then, the aim of this paper that walking through the process of data science like loading data, pre-processing the data and applying various algorithms for analysis were achieved.

At first, through the clustering of the pulsar data-set, the difference of results were caused by the algorithm which was the hierarchical or the non-hierarchical method. The reason of the difference of accuracy rates in two methods to compute the distance of clusters, Euclidian distance and Word's method, were made sure by illustrating the structure of the hierarchical tree as dendrograms.

The second data-set, abalone was the imbalanced data and this report dealt with it by using two approach, the oversampling and the undersampling. The oversampling is the approach which increases the amount of the minority data in the imbalanced column by duplicating it. The undersampling is the technique which decrease the amount of the majority data in the imbalanced column by sampling randomly. This report examined the change of the accuracy rate by increasing the percentage of the undersampled data from 10%. The accuracy rate increased logarithmically and it stagnated at around 5000%. Therefore, as far as analyse results on this data-set, imbalanced data seems not to be problem to decrease the accuracy rate.

In addition, the k parameter of KNN, which means the number of nearest neighbours taken account as close data to predict target variable, was optimized and more than 175 seem to be the best value as the k parameter.

At last, the mushroom data-set was consist of categorical data and one-hot encoding was implemented. Three approaches were carried out to handle missing values, deleting missing values, inserting the most frequent value in the same column and handling missing values as another category. From the result of optimizing the k parameter of KNN, the best method to deal with missing value in this data-set was the insertion the most frequent value since two third of missing values could be thought as the most frequent value if the missing values had the same trend compared with the others. The accuracy rate of SVM were same as the optimized KNN, 1.0. The results were cross validated, but 100% of accuracy rate should be doubted as the overfitting generally. Then, making sure whether the result is overfitted or not is the target of further research.

References

[1] M. Keith, A. Jameson, W. Van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. Bates, N. Bhat, M. Burgay, et al. The high time resolution universe pulsar survey–i. system configuration and initial discoveries. Monthly Notices of the Royal Astronomical Society, 409(2):619–627, 2010.

[2] K. Daisuke, H. Ryuji, H. Keisuke, H. Yuji, Techniques of the data analysis to win in the Kaggle competition, Tokyo: Gijyutu Hyouronsya (in Japanese), 2019.

[3] H. Yuzo, Pattern recognition for beginners, Tokyo: Morikita Syuppan corporation (in Japanese), 2012.

[4] N. Etsuji, Introduction to theories of the machine learning, Tokyo: Gijyutu Hyouronsya (in Japanese), 2015.

[5] H. Tadao, T. Hisatoshi The practical analysis with R, Tokyo: Ohmsya (in Japanese), 2016.

[6] I. Kenichiro, U. Naonori, Pattern recognition by the unsupervesed learning, Tokyo: Ohmsha (in Japanese), 2014.

VI. Appendix

This report used libraries of python below.

scipy : calculate distance of clusters.

numpy : calcurate data as numpy array.

pandas : calculate data as pandas data frame.

sklearn : carry out K-mean, PCA, KNN and SVM.

motplotlib : plot graphs.

sys : expand the limitation to recursive.

random : generate randamized value.

About the classification of the abalone data-set, when the classification was implemented by the KNN with 5 as the parameter k, the accuracy rate was almost same compared it with the case that the KNN was carried out with 175 as the parameter k. However, the line of plot was more smooth when the parameter k was 175.

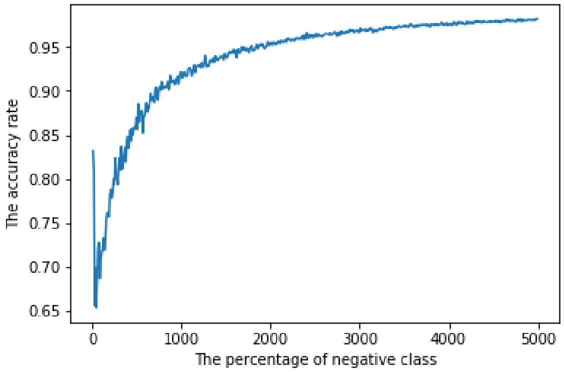

Fig. 10. The percentage of negative values and the accuracy rate on the abalone data-set (KNN with 5 as the parameter k)