K- SUPPORT VECTOR NEAREST NEIGHBOR: CLASSIFICATION METHOD, DATA REDUCTION, AND PERFORMANCE COMPARISON

EKO PRASETYO

Department of Informatics Engineering, Bhayangkara Surabaya University

Jalan Ahmad Yani 114, Surabaya 60231, Indonesia

e-mail: eko@ubhara.ac.id

ABSTRACT

The use of data mining in the past 2 decades in harnessing the data sets become important. This is due to the information given outcome becomes very important, but the big problem are the obstacles data mining task is a very large amount of data. A very large number indeed specificity of data mining in extracting information, but the amount of too big data also cause decrease the performance. On the issue of classification, data that are not positioned on the decision boundary becomes less useful and make classification method is not efficient. K-Nearest Neighbor Support Vector present to answer the problem that data is normally owned by very large data. K-SVNN able to reduce the amount of very large data with good accuracy without degrading performance. Results of performance comparisons with a number of classification method also proves that K-SVNN can provide good accuracy. Among the five comparison methods, K-SVNN got in the big 3 methods. K-SVNN difference accuracy to other methods less of 0.66% on the data set Iris and 20:29% on the data set Wine.

Keywords: *K-Support Vector Nearest Neighbor, classification,performance comparison, Support Vector Machine, ANN Back-Propagation, Decision Tree, Naïve Bayes*

1. INTRODUCTION

The use of management information systems in all fields led to the growth of data increases. In each year, the institution of information system will add a heap data into 'data warehouse' in its database system. In one decade conceivable there will be millions of data remain stored but no gain nothing in the data storage. Data mining is present to utilize data - data such huge numbers. Some things that can be done in data mining are classification, clustering, and association analysis [10], [6].

On the classification task, there is a training process that must be done on the system, before the system is ready to be used for prediction [5]. Although not all methods have the training process as K-Nearest Neighbor, but the method has training process as Artificial Neural Network and Support Vector Machine will require memory resources, and time to do so. In methods that do not have the training process, such as Nearest Neighbor, it requires a large memory resources to store training data.

Basically, training data contains data that is generally in two situations. The first situation is the data in the position around the decisions boundary of classification. In this situation the data have an influence in determining the outcome prediction, this data is required at the prediction time process. In the second situation, the training data that are not positioned on the decisions boundary of classification. The characteristic of these data is in this position is that the data is surrounded by other data that one class with him. These data are not needed at the prediction time. Research conducted attempted to perform data reduction at this position [3]. The process to get the data into a Support Vector by searching for K nearest neighbors of each data. With the choice of the smaller K then the greater data reduction occurs, and vice versa. In his research, the performance accuracy obtained is not different even tend to do better. This method became known as K-Support Vector Nearest Neighbor (K-SVNN).

Research on data training reduction on classification is also conducted [9], the name is bundling text. This method uses a statistical basis in condensing data. In the scheme of proposed method, the data condition on two classes of the most concise is when all data on a class represented by one data point. Condensing data were based on

DOI: https://doi.org/10.54732/jeecs.v1i1.180

statistical average. One data point that condense data represented is obtained by computing average of data that joined that class. The reason for using average because average is also used as base in the Rocchio algorithm and Multinomial Naive Bayes. Application of these methods are still in text data. This method divides the text into several groups of similar data categories. Then calculate average each group to generate a data bundle. The researchers said that bundling text is useful especially when doing text classification using SVM, which the SVM provides high accuracy but many spend time in training process. Thus the method of summarizing data by condensing the entire data into desired data, in this study, data reduction is only performed on data that are not on the decision boundary of classification.

Some other data reduction methods proposed include the Lazarevic and Obradovic [2] in the form of data reduction with Multiple Models Integration. Provost et al. (1999) proposed a reduction by efficient progressive sampling [8]. Freud and Schapire [1] also proposes progressive boosting to reduce data.

The study [4] to test the performance comparison of K-SVNN to other methods, those are Decision Tree (DT), and Naïve Bayes (NB). The results showed that K-SVNN method is better in accuracy, but the training and prediction time are longer. [4] also tested performance comparison of K-SVNN to other methods, those are Support Vector Machine (SVM) and ANN Back-Propagation (ANN-EBP). The test results showed that K-SVNN method relatively better than others. Neither performance accuracy, training and prediction time, K-SVNN method is not necessarily better. This paper presented the results of performance comparative of K-SVNN method when compared to four methods. If in previous studies conducted comparative K-SVNN vs DT and NB, and K-SVNN vs. SVM and ANN-EBP with some choice of parameters. In this study compared performance of K-SVNN vs DT, NB, SVM and ANN-EBP at once. Empirical test with the use of the variation parameter K also conducted to prove the performance more broadly.

The presentation of this paper is divided into 4 sections. Part 1 provides preliminary background of the author doing research. Section 2 presents the research framework. Section 3 presents the testing and analysis were performed. And section 4 presents the conclusions of the research and suggestions for subsequent research.

2. RESEARCH FRAMEWORK

This research observed the K-SVNN method individually as well as when compared to other methods of K-NN, DT, NB, SVM and ANN-EBP. Observations individually performed to test performance on a wide variety of parameters affecting performance. This observation referred to research [3]. While comparisons with other methods conducted to determine how well its performance to other methods on the data with the same test environment conditions.

2.1 Data Reduction

As presented in the previous section, that training data typically contain data that are in two situations. The first situation is the data in position around the decisions boundary of classification. In this situation the data have an influence in determining the outcome prediction, this data is very important and necessary during the prediction process. In the SVM method, this data is great opportunity to be elected as Support Vector, as well as in K-SVNN. In the second situation, the training data is not being positioned on the decision boundary of classification. Characteristic of these data in this position is the data is surrounded by other data that one class with him. In the Nearest Neighbor method, this data is not actually needed at the prediction time. Thus, research conducted [3] attempted to perform data reduction at this position. The process to get the data into a Support Vector by searching for K nearest neighbors of each data. Smaller K selection will result in greater reduction occurs, and vice versa. In his research, the accuracy of the performance obtained is not different even tend to do better. This method became known as K-Nearest Neighbor Support Vector (K-SVNN).

Empirically testing K-SVNN conducted to determine the performance of different K variations. The test results [3] showed that the accuracy is in range 80% even with the use of K ranging from 3 to 15. While the reduction is is done varies from 75% to 40%. The highest reduction can be achieved when K = 3.

In principle, the reduction done is also an impact on the method requires the training process, such as ANN Back-Propagation. During training process, each training data is read to make process of identifying whether accordance with the class or not. If it does not fit then make updates weights. For data that is positioned on the first will give weight change, while in the second position does not give change, but the data through the process of training. Thus, data with this second position needs to be reduced to not perform training process. This data reduction techniques useful to shorten training process because it does not need to process the data that is in the second position, as described [7]. The results of empirical test to determine training time reduction was found that the training time is reduced from 15% to 80%. While accuracy is obtained is also different: the accuracy decreased to 4.76%.

2.2 Performance Comparison

Comparison of five methods are presented in table 1 below. Comparison in the table 1 refer [4] and [5].

Table 1. Comparison of five methods

Criteria	**K-SVNN**	**K-NN**	**SVM**	**ANN-EBP**	**DT**	**NB**
Storing some data train	Yes	Yes	Yes	No	No	No
Criteria affecting performance	K nearest neighbor	K nearest neighbor	alpha, bias, kernel function	number of hidden layers, number of neurons in the hidden layer, learning rate momentum, target error, number of iterations	Criteria for features selection as branch	-
Global optima solution	No	No	Yes	No	No	Yes
Memori space	Large	Large	Middle	Small	Small	Small
Kernel functions utility	No	No	Yes (explicit)	Yes (implisit)	No	No

Table 1 provides overall explanation of results of comparative analysis of five methods. In terms of training data storage partially, K-SVNN have the same principle as the SVM. In terms of the parameters that affect performance, K-SVNN using K as a parameter that determines number of Support Vector obtained, the smaller value of K then less number of Support Vector generated. Global optima solutions are also not obtained by the K-SVNN, because with a different K can be obtained different prediction results too. Memory demands on K-SVNN also great considering the amount of Support Vector generated too much. Actually kernel function only for SVM, SVM has working principles such as map data to a dimension that is relatively higher than its original dimension. While ANN-EBP using hidden layer as a way to map non-linear functions, where the number of neurons in the hidden layer is usually more than neurons in the input layer and output.

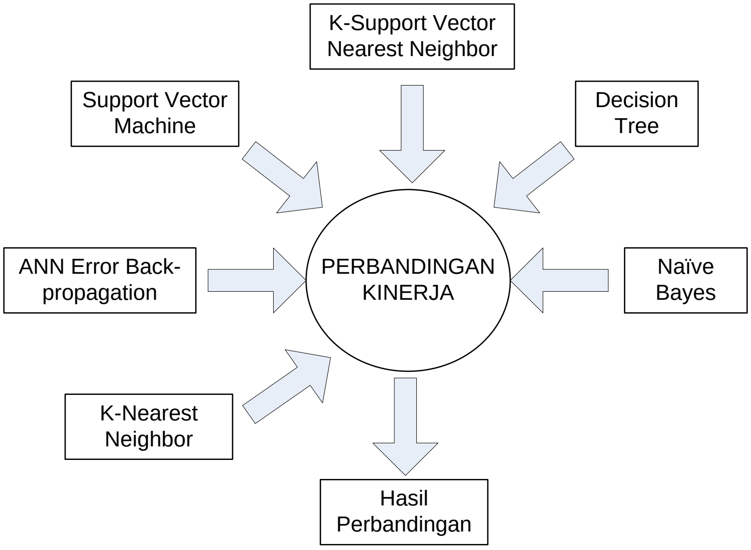

Figure 1. Schematic method comparison

In this study, doing performance comparison of K-SVNN method against other five methods. Comparison diagram is presented in Figure 1. These six methods in its work using different parameters. In this performance test parameters that are used each method selected value that gives the best results. Assumptions such parameters as follows:

1. K-Support Vector Nearest Neighbor
   Observation on the results of research presented [3] provide information that optimal K value used is 9. So, on this performance test used K = 9.

2. K–Nearest Neighbor
   K values used in this study same with K-SVNN that is 9.
3. Support Vector Machine
   The parameters that affect performance of SVM is kernel function. From all types of available kernels, common kernel is RBF. Thus, the performance test is also used RBF.
4. ANN Back-Propagation
   For ANN-EBP, parameters used are as follows: number of hidden layer = 1, number of neurons in hidden layer = 150, learning rate = 0.1, momentum = 0.95, target error = 0.001, and maximum number of iterations = 1000
5. Decision Tree
   Induction method used is C4.5.
6. Naïve Bayes
   There are no specific parameters used.

3. RESULT AND ANALYSIS

3.1 Data set testing

Testing conducted on five public data sets provided by UCI Machine Learning Repository (UCI Machine Learning Repository, [5]), are: Iris (150 records, 4 feature), Vertebral Column (310 records, 6 features), Wine (178 records, 13 feature), Glass (214 records, 9 features), and Diabetic Retinopathy (1151 records, 18 features). Especially for Diabetic Retionopathy data sets, the first feature, not used for quality assessment becase value of the data variance same to 0, the feature can not be used as feature especially Naïve Bayes classification. This caused the value of variant in one of the classes is zero. System testing using 10-fold, of which 90% is used as training data and 10% used as testing data.

Performance testing of K-SVNN in this study is done in two classes only, so it must do merging several different classes into one class on a data set that composition of class more than two, which is Iris. In these data sets, data with the label grade 'setosa' and 'versicolor' combined into one class. Because the data on each feature has range of different values, then the pre-processing is normalization. All data on each feature would be normalized so the value on each feature uses same range of [0,1]. Then do the testing process using these methods.

3.2 Testing Result and Analysis

Testing results of performance accuracy presented in Table 2. From this table, it can be observed that the K-NN method has an accuracy similar to K-NN, and still is positioned on top 3, K-SVNN, K-NN and SVM. In the Grade K-SVNN column, given an * to mark number of other methods of accuracy under K-SVNN.

Table 2. Comparison of performance accuracy

Data set	Accuracy (%)						Grade K-SVNN	Difference to highest accuracy (%)
	K-SVNN	K-NN	SVM	ANN-EBP	DT	NB		
Iris	94.67	95.33	95.33	58.67	61.33	32.00	***	0.66
Vertebral Column	77.10	78.06	83.87	59.68	20.32	27.74	***	6.77
Glass	89.31	89.76	90.74	60.15	76.62	71.49	***	1.43
Wine	75.23	95.52	84.15	53.46	80.13	59.51	**	20.29
Diabetic Retinopathy	63.60	64.90	66.30	54.12	20.33	44.75	***	2.7

When compared to the difference in accuracy between K-SVNN with other methods that accuracy is the highest, K-SVNN provide the difference in the smallest 0.66% on data set Iris (the difference between K-SVNN with SVM or K-NN), and provides biggest difference 20.29% on data set Wine (the difference between K-SVNN with K-NN).

This result indicates that the accuracy prediction given K-SVNN persist well as K-NN and SVM. These test results are good too just limited to the data sets that have been is done performance testing.

4. CONCLUSIONS

From the research conducted, it can be concluded as follows:

1. Method K-SVNN have good performance accuracy as K-NN and SVM method.
2. Method K-SVNN able to perform data reduction training to reduce training computing while maintaining accuracy performance.

The research that has been done has several suggestions for subsequent research as follows:

1. Comparing with other methods of data reduction is also necessary to know the quality of K-SVNN performance with a similar method.
2. It needs further study in proving the performance accuracy on the other data set.
3. It should be proof by the kernel function when compared with Support Vector Machine.

REFERENCES

[1] Freund, Y., and Schapire, R. E., (1996), Experiments with a New Boosting Algorithm, in Proceeding of the 13th International Conference on Machine Learning, Bari, Italy, 325-332

[2] Lazarevic, A., and Obradovic, Z., (2001). Data Reduction Using Multiple Models Integration , In Proceeding of Principles of Data Mining and Knowledge Discovery, Freiburg, 301-313.

[3] Prasetyo, E., (2012). *K-Support Vector Nearest Neighbor Untuk Klasifikasi Berbasis K-NN*, in Proceeding of Seminar Nasional Sistem Informasi Indonesia, Institut Teknologi Sepuluh Nopember, Surabaya.

[4] Prasetyo, E., R.A.D. Rahajoe, S. Agustin, A. Arizal, (2013). *Uji Kinerja dan Analisis K-Support Vector Nearest Neighbor Terhadap Decision Tree dan Naive Bayes*, Eksplora Informatika, 3(1), 1-6.

[5] Prasetyo, E., (2014). *Data Mining – Mengolah Data Menjadi Informasi Menggunakan Matlab*, Andi Offset, Yogyakarta.

[6] Prasetyo, E., S. Alim, H. Rosyid, (2014). *Uji Kinerja dan Analisis K-Support Vector Nearest Neighbor dengan SVM dan ANN Back-Propagation*, in Proceeding Seminar Nasional Teknologi Informasi dan Aplikasinya, Politeknik Negeri Malang, Malang

[7] Prasetyo, E., (2015). *Reduksi Data Latih Dengan K-svnn Sebagai Pemrosesan Awal pada ANN Back-Propagation Untuk Pengurangan Waktu Pelatihan*, SIMETRIS, 6(2), 223-230

[8] Provost, F., Jensen, D., Oates, T., (1999), Efficient Progressive Sampling, in Proceeding of Fifth International Conference On Knowledge Discovery and Data Mining, 23-32.

[9] Shih, L., Rennie, J.D.M., Chang, Y.H., Karger, D.R., (2003). Text Bundling: Statistics-Based Data Reduction, in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC

[10] Tan, P.N., M. Steinbach, V. Kumar, (2006), *Introduction to Data Mining*, 1st Ed, Pearson Education: Boston San Fransisco New York.

[11] UCI Machine Learning Repository , 1 Juni 2014, http://archive.ics.uci.edu/ml/datasets.html