RISK ANALYSIS ON PILE FOUNDATION OF HIGH RISE BUILDING IN SURABAYA

[1]MOHAMAD F.N AULADY, [2]FELICIA T. NUCIFERANI [12]Departement of Civil Engineering,

Technology Institute Adhi Tama Surabaya Jl. Arief Rachman Hakim No. 100 Surabaya

e-mail: [1]mohamadaulady@itats.ac.id, [2]nuciferani@gmail.com

ABSTRACT

Project is a dynamic activity; it will be a lot of risk that will occur in a project due to its dynamic nature. If the risk is not well controlled, it wf32ill lead incompatibility to project specifications, in terms of cost, quality, and time. Therefore, it needs risk management. The application of risk management on construction project has been done considerably. However, most of managed risk are still general. The aim of this research is to specifically review the risks involved in a building, particularly high rise building construction in Surabaya which has pile foundation. Risk identification performed in this research is to determine the level of risk probability and impact on cost performance. The results of the first stage questionnaire were analyzed by the risk matrix method to determine the risk rating. Furthermore, this research also attempts to provide risk responses to the experts regarding of down structure field. The result of the second stage questionnaire were analyzed to decide the risk response policy. The high risk factor on pile construction are the disrepair of retaining wall and the large of foundation shifting, one of the causes is the changing of soil characteristic around the foundation with the type of A2 response, it means that the project is accepted and the risk is controlled by the careful planning.

Keywords: *Foundation, Risk identification, Risk response, Risk Management, Pile*

1. INTRODUCTION

Project is an activity that has a certain period of time with limited resources allocation [1].Hence, project is a dynamic activity, the dynamicity will lead uncertainty that causes the risk in project. Every project has different risk of each activity. Therefore, Risk management has an important role on the construction building project. Construction risks in general are phenomena that affect the project objectives of cost, time and quality. Each stage of the project is related to the various risks and uncertainties that affect both quality and quantity [2]. Risk management is managing the risk from the start to the completion of the project, started with active risk identification, then assessing those risk level, so that the management priority is obtained. Finally, determining the completion steps in order to minimize the risk as much as possible. Thus, it is necessary to do the risk identification for each activity of the project.

Risk management on building construction projects has already done considerably. However, it still refers to general aspects or general risk on the project. Whereas, every project has different activity that cause the general risk could not be applied in the specific project. The building construction project done in Surabaya city are mostly a mall and apartment projects, the mall and apartment construction project can be regarded as a high risk project considering the workload quality and the high structure that will be constructed. The construction process of this project takes long period with high complexity, causing to various risks, starting from the general risk to the specific risk. In the construction project, the very influential working stage on the deviation of project goal is the foundation work. The foundation is the lowest structural part of the building that passes building load to the ground or the rocks located underneath. The pile foundation is used to carry the building load located above water or soft soil to the ground, in order to pass the load to a relatively soft soil until a certain depth so that the building foundation is capable of providing sufficient support to the load [3]. If it is not properly managed, the foundation work is very

risky to the project goal, in terms of cost, quality, and time. The foundation work is also the work that is done earliest. If this foundation work experience the delay process. The delay could also be happening due to inability to manage the risk. In the other hand, the risks that may appear in the foundation work is cost expansion.

According to the background described above, the aim of this research is to determine various risk and risk response regarded as high risk on foundation work in Surabaya. This research is focus on the risk identification based on the contractor point of view and the response done toward risk factor regarded as high risk. This research expect that the practitioners could determine the risk on foundation work before construct the building. Thus, the risk can be reduced. It is also expected that the practitioners can responses the risk properly based on the expert experience.

2. METHODOLOGY

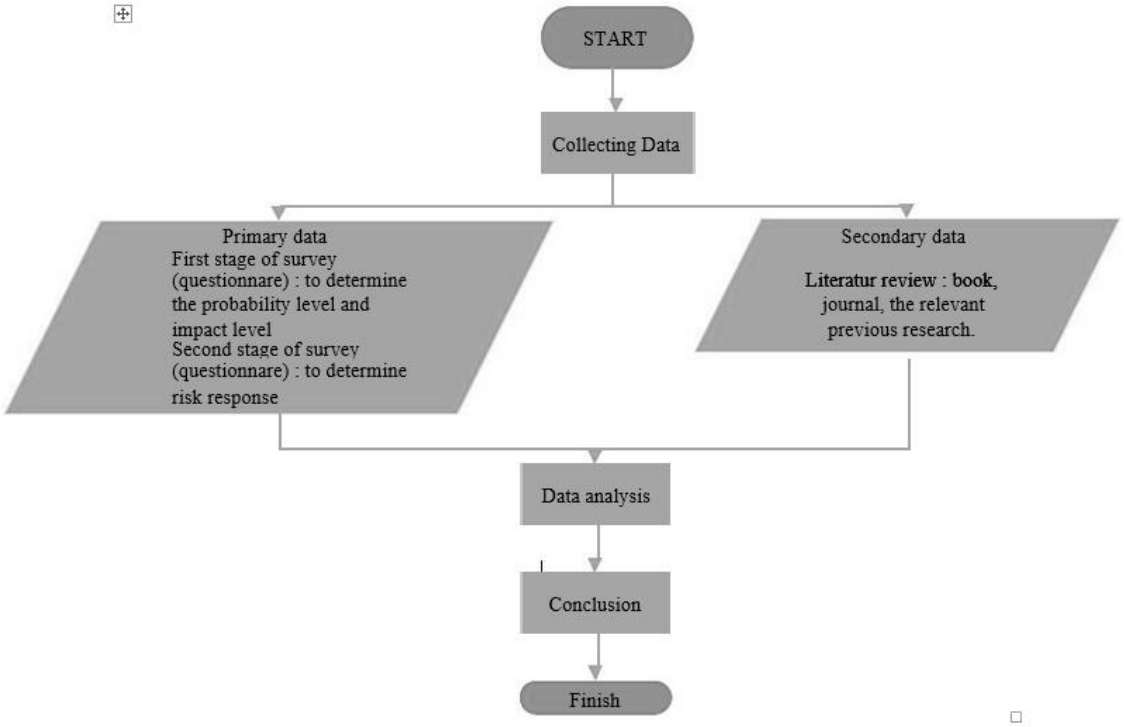

Figure 1. General Flowchart for this research

2.1. Collect Data

Data collecting is done by spreading the questionnaire to the determined project as the research location. Data collecting is done to determine the information needed in obtaining the research purpose, the purpose is showed in hypothesis form that is temporary response toward research question. So that, the response still need to be examined empirically. Thus, it needs data processing. According to the scope of the research, the analysis is only done on Construction project of Tujungan plaza 6 Surabaya, Marvel City Apartment, Gunawangsa Tidar Apartment, Ciputra World Soho and Apartment. The number of respondent in this research is determined by using purposive sampling technique. The purposive sampling technique is determination sampling technique by certain consideration or selection of subject group adjusted with the criteria based on the research purpose. Based on the sampling technique, 3 respondents are determined from each project: Project Manager, Site Manager, and field executive. Because, the three determined respondents have many experience in implementing project. The data used in this research are as follows:

1. Primary data
   Primary data were collected from respondents by directly distributing questionnaires to the projects that were or have been done the soil research, retaining and foundation work. The form of questionnaires used in this study is a stratified questionnaire, stratified questionnaire is the respondent's answer accompanied by a stratified statement. which shows the scale of attitudes covering the range from the

lowest to the highest towards the statement. Respondents in the research are Project Manager, Site Manager and field Executive. The method used to process the data obtained from the questionnaire is the risk matrix method.

2. Secondary Data
   Secondary data is a data or information generated from literature studies, such as: books, journals, and relevant previous studies.

2.2. Data Analysis

The data analysis in this research done by descriptive quantitative survey, the survey is specified by the scale on respondent response. The measurement of respondent perception is done by using Likert Scale then it is processed to determine risk identification.

2.2.1. Risk Identification

Based on the processed data, risk identification is obtained as the following table.

Table 1. Risk Identification for pile foundation

Code	Risk factor	Source
X13	The failure for determining foundation assembly point	[4]
X14	Waiting period until the concrete is ready to use for long-term period	[4]
X15	If it requires cutting in its implementation, it will be difficult and will take a long period	[4]
X16	The disrepair of the retaining wall	[5]
X17	The previous work is late	[5]
X18	The fault of foundation design	[6]
X19	The Fault of determining the foundation dimension	[6]
X20	The cost is high	[4]
X21	The foundation shifting is large	[4]

2.2.2. Risk Analysis

Measuring respondent perception cannot be directly processed because the value is still qualitative, so it must be quantified by giving a scale on respondent answer, by specifying the code to facilitate the data processing mathematically [7]. There are two types of risk analysis, as follows:

1. Qualitative risk analysis
   Qualitative data can simply be interpreted as not number data. In qualitative data, the data cannot be performed as mathematical operations [5] Qualitative data can be divided into two, nominal and ordinal data. The nominal data is the lowest data in the data measurement level. While ordinal data is higher than the nominal data.
2. Quantitative Risk Analysis
   Quantitative data can be interpreted as the form of numbers. Thus, various mathematical operations can be performed on quantitative data [5].

In this research we first conduct quantitative risk analysis to determine the most dominant risk. Each expert is given a question based on table 1. The experts will give a value of 1-5 for how much risk

factors that may occur and also how much impact is on the same scale. Then the expert answer is averaged with the formulation below:

$$Index\ value = \frac{(F1x1)+(F2x2)+(F3x3)+(F4x4)+(F5x5)}{5} \quad (1)$$

Which:

F1 = respondents answer frequency that answer scale 1

F2 = respondents answer frequency that answer scale 2

F3 = respondents answer frequency that answer scale 3

F4 = respondents answer frequency that answer scale 4

F5 = respondents answer frequency that answer scale 5

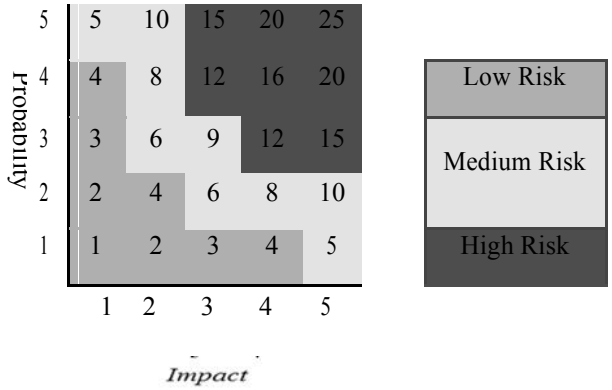

Figure 2. Matrix table of risk category

After this process, we will obtain probability and impact value in each risk factor. It will be compared to figure.2 to determine risk category.

The entry process of matrix table is done by substituting the result of risk index formula to determine probability and impact scale. The result of risk index formula is analyzed using matrix table with *Probability x Impact* (P x I) formula.

After substituting those results, it is continued by multiplying the scale on the probability and impact column as follows:

Table 2. Risk analysis based on impact toward the cost

CODE	P	*I*	P x *I*	Risk Categories
X13	2.6	3.1	8.06	Medium
X14	2.8	2.6	7.28	Medium
X15	2.6	3.8	9.88	Medium
X16	5	5	25	High
X17	3	3.2	9.6	Medium

X18	2.8	3	8.4	Medium
X19	2.6	3.8	9.88	Medium
X20	3.2	2.5	9	Medium
X21	3.6	4.8	17.28	High

Sources: Processing data (2016)

The result of multiplying *Probability x Impact* (*P x I*) on Table 2 become the reference to determine risk factor that may cause the dominant impact toward the cost. Based on Table.2 we will obtain 2 risk categories that is dominant enough, that are the disrepair of the retaining wall and the large of foundation shifting. Furthermore, in the response risk, we will focus on the most dominant risk factor.

2.2.3. Risk Response

To determine risk response, qualitative risk analysis were done by giving the expert question. The experts is interviewed based on the most dominant risk factor. Then, they will answer how to reduce the risk based on their experience. Based on risk factors obtained through analysis the risk factor regarded as high risk rating, questioner's distribution of stage 2 is done to determine the reason factor and to determine the response toward risk factor. Here is the response policy toward risk factor according to [7]:

1. The Project is rejected (T1).
2. The Project is accepted, but the risk is certified by the owner (T2).
3. The Project is accepted, and the risk is redirected to other parties in the company guided (A1).
4. The Project is accepted, and the risk is controlled by company it is self with the careful planning (A2).
5. The Project is accepted, and the risk is accepted as the cost, it means that if the risk occurs, it has been calculated in budget (A3).

3. RESULT

Based on the processing data, it can be obtained that the very dominant risk factor is the disrepair of retaining wall (X16). Therefore, it is necessary to do the risk response to determine how to anticipate by the expert experience. Table 3 will explain how to do the risk identification toward the response.

Table 3. Risk Response for high risk

Risk Identification	Cause Analysis	Responce Policy					Reponse
		T1	T2	A1	A2	A3	
The desrepair of retaining wall (X16)	The soil stability is poor				√		Accurately investigating the soil first.
	Earthquake					√	Predicting the earthquake level and build the cooperation with BMKG.
	Soil data is not accurate			√			1. Turn over the occurred risk to delegated sub-contractor. 2. monitoring the soil while investigation process.
	Error calculation of horizontal force in the soil			√			Turn over the occured risk to specialist sub-contractor.
	The fault of geogrid installation			√			Turn over the geodrid installation to specialist sub-contractor.
	The sloping surface due to Compacation using to heavy equipment			√			Turn over the work to the specialist subcontractor
	Design fault of SPT				√		Periodically monitoeringmust be done on planing/ design stage

	Design fault of retaining wall			√			1. Redesign. 2. Trun over the work to the expert.
	The fault of geogrid utilization				√		The gegroid installation planing must be ajusted with field needs.
	Load –carrying capasity is poor				√		1. Soil reparation by geotextile installation. 2. Properly investigating the soil in order to accurately determining the real load-carryingcapasity.
The large of foundation shifting (X 21)	Flood					√	1. Control tub installation for flood water reservoirs. 2. Must have BMKG estimation data.
	The changing of soil characteristic around the foundation				√		Accurately invertigating the soil.
	The soil stability is poor				√		The repairation of soil stability by the variuos of geotextile as requirement.
	DPT is poor				√		1. Protection installation in the form of sheet pile. 2. Monitoring periodically. 3. The construction method that has been discussed with the expert.
	Load-carrying capasity is poor				√		1. Soil repairation by geotextile installation. 2. Properly investigating the soil in order to accurately determining the real load-carrying capacity.
Risk identification	The huge earthquake that cause the foundation shifting					√	1. Strengthen foundation design to reduce the foundation shifting. 2. Must have eartquake estimation data from BMKG.

Table 3 explain what must be done by practitioners based on expert experience and response once the risk occur. It can be known that when descrepair for retaining wall occur due to the fault of geogrid utilization, the expert still accept the project and the risk is controlled by company its self with the careful planning (A2 response), the expert suggest the company to do geogrid installation planing and it was adjusted with field needs. Another reason for derepair of retaining wall is design fault of retaining wall, in this case, the expert still accept the project, however the risk is redirected to other parties in the company guided (A1 response). And the expert suggest the company to do redesign of retaining wall and turn over the occurred risk to expert.

4. CONCLUSION

The conclusion of the research is:

Response type of risk factor regarded as high risk is determined by last stage questionnaire distribution. Research variables regarded as high risk have the same probability and impact multiplication. One of them is pile foundation project with research variable of the disrepair of retaining wall that have several cause analyses, they are: the poor soil stability, design fault of SPT by response type A2 which means the project is accepted and the risk is controlled by the careful planning.

Most all the risk policy is A1, A2, and A3. It means that the experts agreed to accept the project, and the risk is redirected, controlled and accepted with additional cost based on the work type.

References

[1] Y. A. Messah dan P. Soekirno, "The Suitability Study of Construction Service Law No 18 of 1999 and The Presidential Decree No 80 of 2003 About Goods and Service Procurement of Government Construction Works," Institut Teknologi Bandung, Bandung, 2009.

[2] Nurlela dan H. Suprapto, "Risk Management Analysis and Identification of Construction Project of High Rise Building Infrastructure," *Construction Design Journal,* vol. 13, no. 2, pp. 114-124, December 2014.

[3] H. C. Hardiyanto, Foundation Planning and Analysis Part 1, Yogyakarta: Gadjah Mada University Press, 2011.

[4] Marwan, "Supporting Capacity of Pile Foundation on Religious Strait Court building in Long Strait of Meranti Regency," Islamic University of Riau, Pekanbaru, 2011.

[5] G. R. Maharani, "Time and Cost Risk Management on Structure Work of High Rise Building Project in Jabodetabek," Universitas Indonesia, Depok, 2011.

[6] A. D. Iriani, Land and Foundation Work Risk Analysis, Depok: Universitas Indonesia Press, 2008.

[7] Asiyanto, Risk Management for Construction Project, Jakarta: Pradnya Paramita, 2008.

[8] H. M. Tumimor, "Risk analysis on Brigde Construction in North Sulawesi," *Dimensi,* vol. 6, no. 2, pp. 235-241, 2014.

CLASTERING TYPE OF BEST GRAMEDIA PUBLISHER SELLER PAPER USING SHRINKING SHARED NEAREST NEIGHBORS METHOD

[1]NURVIDI RATNA SARI, [2]RIFKI FAHRIZAL ZAINAL, ST.,M.Kom.,

[3]RANI PURBANINGTYAS, S.Kom., MT.

[123]Informatics Engineering Study Program, Informatics Engineering, Bhayangkara University

Jl. Ahmad Yani 114, Surabaya, East Java, 60231

e-mail: [1]vidisoedarsono@gmail.ac.id, [2]rifky@ubhara.ac.id, [3]rani@ubhara.ac.ic

ABSTRACT

Over time in the business world of book publishing has increased the number of publishers is very significant. This has triggered a very tight competition to grab the attention of customer for the next operational activity. The publisher must be able to guess which type of book the customer candidate will be interested in. Given these problems, the publishing industry is still having trouble determining what kind of books will attract potential readers. So this situation hampers publishers to be able to create the kind of book that the readers want. Then to make it easier to determine and know the best seller type book group is the author uses one way that is by using SSNN method.

SNN-based data shrinking algorithm (SSNN) uses the concept of data movement from the shrinking data algorithm to increase the accuracy obtained. The concept of data movement will strengthen the density of adjacent graph so that cluster forming process can be done from neighboring graph components that still have adjacency relationship. In this trial using data order gramedia property and using 1200 test data which is divided into 4 groups of test data. Produces clusters that generate the amount of data per genre on each cluster. And generate the children's book genre as a genre of best seller books.

Keywords: book, SSNN, best seller, genre

1. INTRODUCTION

Competition to grab the attention of customer for the operation of book publishing operational very significant lately. Thus to make it easier to determine and know the type of best seller book is the author uses one way is to use the SSNN method. The Nearest Neighbor (SNN) Shared algorithm is an algorithm that forms a graph of consideration that uses the similarity between data points based on the number of closest neighbors held jointly. The cluster is obtained from the representative points selected from the adjacent graph. The representative point is used to reduce the number of clustering errors, but also reduce accuracy. SNN-based data shrinking algorithm (SSNN) uses the concept of data movement of the shrinking data algorithm to increase the accuracy obtained. The concept of data movement will strengthen the density of adjacent graph so that cluster forming process can be done from neighboring graph components that still have adjacency relationship. Thus can be determined what type or genre wanted by customer.

2. METHODOLOGY

2.1 Data Mining

Data mining is a process that uses Static, Mathematical, Artificial intelligence, and machine learning techniques to extract and identify useful data information and related knowledge from large databases.[1]

Data mining is a series of processes to explore the added value of a data set of knowledge that has not been known manually. In the big journals of data mining is also called Knowledge Discovery In Database (KDD) is an activity that includes the collection, use of historical data to find regularities, patterns or relationships in large data sets.[2]

2.2 Clustering

Clustering is a grouping method based on the size of proximity (resemblance). Clustering is different from group, if group means group same condition if not yes definitely not group. But if the cluster does not have to be the same but the grouping is based on the proximity of an existing sample characteristic, one of them by using the euclidean distance formula. The application of this cluster is very much, because almost in identifying problems or decision-making is always not exactly the same but tend to have similarities only.[3]

2.3 ***Algorithm Shrinking based Shared Nearest Neighbor (SSNN)***

SSNN algorithm is a development of SNN algorithm. The cluster accuracy is improved by increasing the density of the graph of the customer formed in the SNN algorithm. Then to strengthen the density is to use the concept of data movement toward the cluster center of the data algorithm Shrinking.

Because, this repair algorithm will make the data points seem to shrink toward the cluster center, then this algorithm is named SNN algorithm based data shrinking or *Shrinking based Shared Nearest Neighbor (SSNN)*. [3]

If the data points in the neighboring graph are moved toward the cluster center, the neighboring weights for the data points in the same cluster will become larger and the neighboring weights for the data points in the different clusters will become smaller. To be able to apply the concept, the neighboring graph will be formed in several iterations. The number of nearest neighbors used for the formation of an adjacent graph increases in each iteration. Overall, the SSNN algorithm consists of steps:

1. Calculate the similarity value of the data set.
   To calculate the value of similarity (distance):
   Formula :

$$\text{Calculating distance} = \sqrt{\sum(x - x_1)^2}$$

explanation :
x : First set of data sets.
x_1 : The next set of data sets.

2. Form a list of the nearest neighbor k each data point
3. Form adjacent graph from the list of nearest neighbor's k
4. Calculate the value of proximity and disconnect the adjacency weight less than the closeness value
5. Repeat steps 3 and 4 until the closeness value is greater than the proximity threshold value6. The cluster form of the adjacent graph component still has the weight of neighboring relationships.

Steps 1 to 3 are still the same as the SNN algorithm, but in step 2 the list of nearest neighbors is formed for the entire data point in the data set because it is required for the establishment of an adjacency graph in several iterations. Steps 3 and 4 are done in several iterations to apply the concept of data point movement toward the cluster center.

The SSNN algorithm enters the sequence of the nearest neighbor in the list of nearest neighors in the calculation of the weight of the neighboring relationship so that the value of the neighboring weights will have a more accurate value.

For example, there are two data points i and j, then the neighboring weights of i and j are:

$$\text{weight } (i, j) = \sum(k+1-m) \times (k+1-n)$$

explanation :
K = The size of the list of nearest neighbors
m and n = The position of the nearest neighbor listed in the nearest neighbor list i and j.

The k value will increase in each iteration according to *the Move Points (MP)* parameter to be able to adjust the change of adjacent graph. The density value of each data point is calculated based on the total weight of the neighbor relationships owned by each data point.

Data points in the same cluster have adjacent density values. The SSNN algorithm uses the condition as a proximity value to determine the weight of the adjacent neighbor relationship to be disconnected and ignored in the adjacency graph formation process in the next iteration. In other words, the weight of neighbor relationships that are not close to their density values will be disconnected and ignored. The process of calculating the value of proximity in each iteration serves to overcome the shortcomings of the basic SNN algorithm that is too dependent on a threshold value of the neighboring weights.

The iteration process will be stopped if the proximity value has reached a closest or nearest neighbor threshold value that is used is greater than the amount of data in the data set. The restrictions are made to prevent the occurrence of cluster splits.

The neighboring graph obtained at the end of the iteration will have a strong density so there is no need for a process of determining representative points. The cluster can be directly formed from an adjacent graph component that still has the neighboring weights. Related data points will be included in the same cluster, so cluster forming process can be done in a shorter computation time.

3. RESULTS

3.1 Test Results 1

The first test used the top 40 data from the first quarterly test data. Generates 10 Clusters that generate 1124 data of nearest neighbor data. This first test shows that the genre of Children Books is superior among other genres of books. This can be seen in Table 6.9 where in the test each cluster has a varying number, the test that produces 10 clusters shows the first cluster is the most superior among other clusters. In the first cluster we can see that the owner of the largest amount of data is Children's Book. In addition, when viewed from the total number of genres in the 1st trial that has the highest amount of data Children's Book with calculations Children's Book 759 data, Cooking 33 data, Diet & Health 33 data, Economi & Bussines 35 data, Enterteinment 33, Fiction & Literature 99 data, Reference & Dictionary 33 data, Self-Improvement 66 data and Social Science 33 data. We can thus conclude that in experiment 1 it generates the Children's Book genre as the most genre of books in interest or best seller book genre.

Tabel 1. Results of testing genres 1

Genre / Cluster	Children's Books	Cooking	Diet & Health	Economi & Bussines	Enterteinment	Fiction & Literature	Reference & Dictionary	Self - Improvement	Social Sciences	Total
1	144	3	2	0	3	8	3	5	2	170
2	82	2	1	35	8	5	8	2	1	144
3	108	2	2	0	3	20	3	4	2	144
4	55	3	28	0	3	34	3	28	2	156
5	103	3	0	0	3	6	3	6	26	150
6	72	2	0	0	3	5	3	21	0	106
7	76	2	0	0	3	4	3	0	0	88
8	62	16	0	0	3	6	3	0	0	90
9	27	0	0	0	2	11	2	0	0	42
10	30	0	0	0	2	0	2	0	0	34
Total	**759**	**33**	**33**	**35**	**33**	**99**	**33**	**66**	**33**	1124
Total Data	**1124**									

3.2 Test Results 2

In test 2 it still uses the top 40 results in quarter 2. The result of trial 2 yields 10 clusters, the number of clusters is almost the same as the test result 1 but has the smallest number of neighboring data is smaller that is 942 data. Of the 10 clusters retrieved trial 2 still has the best cluster of cluster 1. In cluster 1 has the largest amount of data than the other clusters. In the cluster again shows that the genre of Children Books is the most superior among other genres of books. In addition, the results of the total number of clusters of each genre also show the genre of Children's Book which has the highest number of Children's Book 637 data, Cooking 63 data, Diet & Health 60 data, Economi & Bussines 30 data, Fiction & Literature 60 data, Medical 32, Religion & Spirituality 30 data and Social Science 30 data. Thus it can be concluded again that the genre of Children Book is a genre of bestsellers. More detailed calculations can be seen in table 2.

Tabel 2 Results of testing genres 2

Genre / Cluster	Children's Books	Cooking	Diet & Health	Economi & Bussines	Fiction & Literature	medical	religion & spirituality	Social Sciences	Total
1	98	5	4	4	26	0	4	4	145
2	65	2	3	2	3	1	1	1	78
3	20	32	2	1	2	1	1	1	60
4	47	1	9	8	9	30	11	11	126
5	48	1	2	1	2	0	1	1	56
6	86	2	4	2	4	0	2	2	102
7	67	2	27	3	5	0	2	2	108
8	78	2	3	3	3	0	2	2	93
9	84	3	3	3	3	0	3	3	102
10	44	13	3	3	3	0	3	3	72
Total	**637**	**63**	**60**	**30**	**60**	**32**	**30**	**30**	**942**
Total Data	**942**								

3.3 Test Results 3

The third test uses the top 40 data from the 3th quarterly test data. Test 3 produces 8 clusters, fewer clusters than the previous test, test 3 produces 1264 data of nearest neighbors. Of the 8 clusters obtained in trial 3, the first cluster is the most superior among the other clusters, because it has the highest amount of data among other clusters. In the genre of cluster 1, the Children's Book genre has the largest amount of data among other genres. Not only that, if calculated as a whole in test 3 or quarter 3 this genre of Children's Book again has the most amount of data than the other genres: Children's Book 875 data, Diet & health 35 data, Economi & Bussines 70 data, Enterteinment 38 data, Fiction & Literature 106 data, Self- Improvement 70 data and Social Sciences 70 data. Thus it can be concluded that the genre of Children's Book back into the genre of best seller books in this trial. More detailed test results can be seen in table 3.

Tabel 3 Results of testing genres 3

Genre / Cluster	Children's Books	Diet & Health	Economi & Bussines	Entertein ment	Fiction & Literature	Self – Improvement	Social Sciences	Total
1	218	5	13	0	38	46	60	380
2	67	2	5	38	40	4	4	160
3	195	3	7	0	7	6	6	224
4	40	25	27	0	2	4	0	98
5	195	0	11	0	4	6	0	216
6	70	0	2	0	2	4	0	78
7	61	0	3	0	2	0	0	66
8	29	0	2	0	11	0	0	42
Total	**875**	**35**	**70**	**38**	**106**	**70**	**70**	**1264**
Total Data	**1264**							

3.4 Test Results 4

The fourth test uses the top 40 data from the 4th quarterly test data. Test 4 produces 8 clusters, test 3 yields 1272 data of nearest neighbors. Of the 8 clusters obtained in pilot 4, the first cluster is the most superior among the other clusters because it has the highest amount of data among other clusters. In cluster 1, there is a considerable composition of the Children's Book genre in the genre of the others. Then if the total count is calculated in 4, the Children's Book genre has a large amount of data in the amount of other genres. Thus it can be concluded that the genre of Children's Book is a bestseller genre or best seller on trial 4. Detailed test results are found in table 4.

Tabel 4 Results of testing genres 4

Genre / Cluster	Children's Books	Cooking	Economi & Bussines	Fiction & Literature	Medical	Reference & Dictionary	Self - Improvement	Total
1	339	55	6	5	6	13	18	442
2	203	5	4	2	4	5	7	230
3	136	4	20	2	2	4	6	174
4	55	3	3	26	21	3	29	140
5	84	2	2	0	2	2	2	94
6	104	4	0	0	0	4	4	116
7	36	2	0	0	0	2	2	42
8	28	2	0	0	0	2	2	34
Total	**985**	**77**	**35**	**35**	**35**	**35**	**70**	**1272**
Total Data	**1272**							

4. Conclusions

From the results of research that has been done can be concluded as follows:

1. The use of this method is quite easy because to normalize we can determine its own requirements.
2. SSNN method can be used to cluster genre on best seller book determination.
3. The results of this experiment indicate that in this period the genre of Children book is a genre of bestseller or most desirable.

REFERENCE

[1] Kusrini and Lutfi, E.T (2009), *Algoritma Data Mining*. Yogyakarta: Andi Offset.

[2] Pramudiono (2006), *Apa Itu Datamining?*

[online], (http://gunawandra.blogspot.com/2013/03/pengertian-data-mining-menurut-para.html.)

[3] Zainal, RF dan Djunaidi A (2008), *Algoritma Shared Nearest Neighbor Berbasis Data Shrinking.* Vol .7 No.1, Institut Teknologi Sepuluh November. Surabaya.

RELATIONSHIP BETWEEN DATA REDUCTION AND PERFORMANCE IMPROVEMENT OF CLASSICATION WITH K-SUPPORT VECTOR NEAREST NEIGHBOR

EKO PRASETYO

Department of Informatics Engineering, University of Bhayangkara Surabaya

Jl. Ahmad Yani 114 Surabaya

e-mail: eko@ubhara.ac.id

ABSTRACT

K-Support Vector Nearest Neighbor (K-SVNN) as a Nearest Neighbor-based method can be used for data reduction. Reduction is done by giving parameter K as nearest neighbor used. Besides data reduction, K-SVNN can also improve the performance of prediction accuracy. The author tests the value of K that is the percentage of the amount of data. The K used varies from 10% to 100%. This study was conducted to observe the relationship between data reduction and performance improvement. Performance improvement is measured by Fisher Discriminant Ratio (FDR). From the results of research proved that in some data sets, data reduction with K-SVNN can increase the FDR value, while some other data sets can't. The K value that gives the highest FDR value is 30% to 50% with data reduction up to 4.49%.

Keywords: *K-Support Vector Nearest Neighbor, data reduction, performance, improvement, classification, Fisher's Discriminant Ratio*

1. INTRODUCTION

K-Support Vector Nearest Neighbor (K-SVNN) is a method based on Nearest Neighbor to perform classification task [1]. This method successfully reduces the training data used during the training process but still strives to maintain accuracy. The results of the experiments performed show that the yielded reduction performed for K = 7 is up to more than 49% but the accuracy obtained can still survive, even raise 1% compared to other methods with the highest accuracy. One of the expected goals of K-SVNN is that it can reduce the prediction time used, as given by Prasetyo [1]

The comparison of K-SVNN with Decision Tree (DT) and Naïve Bayes (NB), in terms of time spent on training and prediction, K-SVNN is longer than DT or NB [2]. This is because K-SVNN must calculate the distance between each data with all other data, whereas DT and NB do not do that but other work where consumption time is less. The result of comparison of K-SVNN with DT and NB on prediction accuracy was K-SVNN better than DT and NB. This becomes something interesting, because on the one hand, K-SVNN tries to reduce training data but still can keep accuracy obtained. Comparison of other methods of ANN Back-propagation (ANN-EBP) and Support Vector Machine (SVM) have also been done [3]. The comparison results show where the K-SVNN accuracy tends to be superior to ANN-EBP and SVM, although for the Vertebral Column and Wine data sets, K-SVNN is lower than SVM. As for training time, K-SVNN is superior to other methods. For predictive time is shorter than ANN-EBP but longer than SVM. Comparison with four data sets proves that K-SVNN has a relatively higher performance than ANN-EBP, DT and NB methods, but lower than SVM and K-Nearest Neighbor [4]. The research that has used K-SVNN as preprocessing is Prasetyo [5], In his research, K-SVNN was used for data reduction before applying the ANN-EBP classification method. Tests performed on five data sets gave results that the performance on the Wine data set was poor. In this study, the author tries to prove the use of K-SVNN on some data sets with a choice of K that is proportional to number of data.

From the exposure provided by [2, 3], K-SVNN has performance that competes with other classification methods. Sometimes lower and sometimes higher for both accuracy, training time and prediction time. This study performs empirical testing on the effect of data reduction practice with predictive accuracy obtained. Previous

research has done a reduction with certain selected K, [1] use K=7, [2, 3] use K=13. The number of data from the test data sets is not the same, so in this study the authors used the percentage to measure the effect. The results of this study can prove whether K-SVNN can provide benefits of data reduction training and increased accuracy based on certain percentage reduction choices. The data set feature strength measurements in the authors' prediction using Fisher's Discriminant Ratio [6, 7].

The exposure of this paper is divided into four parts. Part 1 provides preliminary background of the author doing research and research related to the study. Section 2 presents the methodology of the study and design of systems. Section 3 presents the results and discussion. And section 4 presents the conclusions of the research and suggestions for next research.

2. METHODOLOGY

2.1 Research Methodology

Research methodology conducted by authors is presented in Figure 1. The input given is training data and K value of K-SVNN. Training data used by authors is part of the data set with a certain percentage. While K is a choice of K that can be set independently by user.

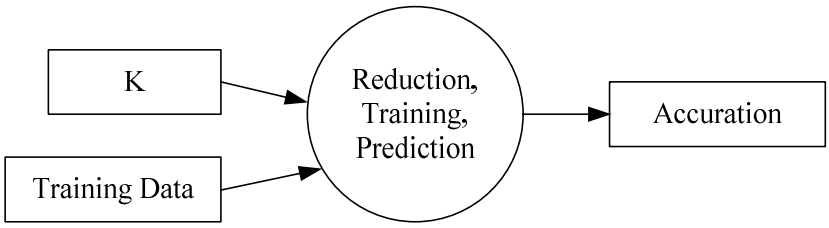

Figure 1. Research methodology

The authors use percentage as K value of K-SVNN. K serves as a determinant of reduction parameter, where K is the number of nearest neighbors involved in K-SVNN calculations. The data sets used by the authors have unequal number of data, then for proportional needs in the reduction authors use value of K with the percentage according to the number of data from the dataset. For example, for K = 10% of the data, then for data set with the number of data 200 will use K = 20, while for the data set with the amount of 500 data will use K = 50. In this study the authors use 10 variations of K, that is 10% , 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. For K = 100% gives a 0% reduction result, so the FDR obtained must be the same as the K-NN method without data reduction.

The result of K-SVNN test with 10 variations of K as the author's way to know the correlation between the amount of reduction made to the prediction accuracy obtained. With reduced amount of training data due to reduction whether it can keep accuracy, decrease or increase. As in previous studies that the reduction done on the train data is able to keep the accuracy of predictions obtained. This study can prove the behavior.

To know the strength of training data of the results of the reduction at the time of prediction, the authors use Fisher's Discriminant Ratio (FDR) [6, 7]. FDR value is calculated on each feature, then calculated the sum of FDR value on each dataset. The sum of FDR value is calculated on the 10 K variations of K-SVNN specified in this study. Furthermore, the authors do the analysis on the results obtained.

3. RESULTS

3.1 Testing Result

The author tests on 5 public datasets downloaded from UCI Machine Learning Repository, they are: Iris (150 records, 4 feature), Vertebral Column (310 records, 6 features), Glass (214 records, 9 features), Wine (178 records, 13 feature), and Diabetic Retinopathy (1151 records, 18 features). Especially for Diabetic Retinopathy data sets, we don't use the first feature because value of the data variance same to 0. This caused the value of variant in one of the classes is zero. System testing using 5-fold, of which 80% is used as training data and 20% used as testing data.

The results of the tests that the author did on the data set Iris presented in Table 1. The reduction rate obtained proportional. It can be observed in the Reduction column, its value decreases as the value of K increases. The FDR value obtained for feature 1 continues to decrease as the value of K decreases. This result indicates that the 1st feature is proportional to the value of K but inversely proportional to the size of the reduction experienced. For features 2, 3 and 4 FDR values rise and fall along the value of K and the reduction experienced. The ability to separate classes can be seen in the Sum of FDR column, where Sum of FDR values are obtained by summing up all

FDR features. But the highest value of Sum of FDR is obtained only when 0% reduction or no reduction is done. This behavior indicates that for Iris data sets it is not suitable for reduction with K-SVNN.

Table 1. Result of FDR from Iris data set

No	K (%)	Reduction (%)	FDR of Feature				Sum of FDR
			1	2	3	4	
1	10	67.33	0.09	0.01	2.33	2.55	4.98
2	20	50.00	0.18	0.02	2.68	3.33	6.22
3	30	42.00	0.37	0.08	3.01	3.79	7.25
4	40	39.33	0.43	0.12	3.24	3.91	7.70
5	50	36.00	0.53	0.16	3.22	4.08	7.99
6	60	34.00	0.61	0.18	3.19	4.20	8.19
7	70	21.33	1.00	0.01	2.11	3.01	6.14
8	80	4.00	1.43	0.03	2.83	3.71	8.00
9	90	0.67	1.51	0.04	2.99	3.86	8.40
10	100	0.00	1.53	0.05	3.01	3.89	**8.48**
11	Without reduction		1.53	0.05	3.01	3.89	**8.48**

The results of the testing that the authors do on the data set Vertebral Column is presented in table 2. The FDR value given for feature 1 is also straight proportional to the decrease in value of K and inversely proportional to the amount of reduction experienced. Features 2 to 6 have FDR values that vary increase and decrease throughout the K value and reduction options experienced. The highest value of Sum of FDR is obtained by K = 50% with a value of 2.61, higher than without reduction of 2.10.

Table 2. Result of FDR from Vertebral Column data set

No.	K (%)	Reduction (%)	FDR of Feature						Sum of FDR
			1	2	3	4	5	6	
1	10	16.77	0.26	0.20	0.18	0.12	0.44	1.08	2.27
2	20	9.68	0.28	0.20	0.21	0.13	0.40	1.20	2.41
3	30	6.45	0.32	0.24	0.24	0.15	0.37	1.22	2.54
4	40	4.52	0.34	0.26	0.25	0.14	0.35	1.20	2.54
5	50	2.58	0.35	0.28	0.26	0.14	0.35	1.23	**2.61**
6	60	1.61	0.35	0.29	0.27	0.13	0.31	1.17	2.52
7	70	0.97	0.36	0.30	0.28	0.12	0.30	1.17	2.53
8	80	0.65	0.36	0.31	0.28	0.12	0.28	1.14	2.50
9	90	0.32	0.36	0.32	0.28	0.12	0.27	1.14	2.49
10	100	0.00	0.36	0.31	0.28	0.12	0.28	0.75	2.10
11	Without reduction		0.36	0.31	0.28	0.12	0.28	0.75	2.10

The results of the testing that the authors do on the data set Glass presented in Table 3. FDR values provided for features 1 and 9 straight proportional to the decrease in value of K. While features 6 and 7 inversely proportional. Features 2, 3, 4, 5, 7, and 8 have varying FDR values increase and decrease throughout the K value and reduction options experienced. The ability to separate classes has the highest Sum of FDR value obtained by K = 30% with a value of 6.40, higher than without a reduction of 5.62.

The results of the testing that the authors do on the data set Wine presented in Table 4. FDR values provided for feature 1 straight proportional to the decrease in value of K. Features 6 to 9 has a very small FDR value, zero in all options K used, this proving that the feature is not informative used in the classification. While other features have FDR values that vary increase and decrease throughout the selection of K value and reduction experienced. The ability to separate classes has the highest Sum of FDR value obtained by K = 40% with a value of 6.94, higher than without reduction of 6.89. This good result is different from the results of the research presented by Prasetyo [5] where the accuracy performance becomes poor when using K-SVNN for data reduction. Higher results in the study are global with FDR value.

Table 3. Result of FDR from Glass data set

No.	K (%)	Reduction (%)	FDR of Feature									Sum
			1	2	3	4	5	6	7	8	9	
1	10	38.79	0.00	0.36	2.64	0.72	0.09	0.10	0.08	0.50	0.06	4.56
2	20	7.94	0.01	0.56	3.60	1.17	0.06	0.09	0.05	0.75	0.09	6.38
3	30	2.34	0.02	0.55	3.83	0.97	0.08	0.09	0.04	0.72	0.10	**6.40**
4	40	1.87	0.02	0.55	3.64	0.98	0.09	0.09	0.03	0.73	0.10	6.22
5	50	1.40	0.03	0.55	3.31	0.98	0.09	0.06	0.01	0.76	0.11	5.90
6	60	1.40	0.03	0.55	3.31	0.98	0.09	0.06	0.01	0.76	0.11	5.90
7	70	0.93	0.03	0.56	3.16	0.96	0.09	0.06	0.01	0.65	0.11	5.63
8	80	0.93	0.03	0.56	3.16	0.96	0.09	0.06	0.01	0.65	0.11	5.63
9	90	0.93	0.03	0.56	3.16	0.96	0.09	0.06	0.01	0.65	0.11	5.63
10	100	0.00	0.05	0.50	3.28	1.01	0.05	0.00	0.00	0.60	0.12	5.62
11	Without reduction		0.05	0.50	3.28	1.01	0.05	0.00	0.00	0.60	0.12	5.62

The results of the testing that the author did on the data set Diabetic Retinopathy presented in Table 5. Actually, the FDR value provided by all the features in the table is low, almost zero all. The Sum of FDR value obtained in the data set without reduction is 0.92, while with reduction, Sum of FDR value varies above it, the highest is 0.98 with K 40%.

The FDR value given for feature 1 is straight proportional to the value of K. Features 6 to 9 has a very small FDR value, zero in all K options used, this proves that the feature is not informative used in the classification. While other features have FDR values that vary increase and decrease throughout the selection of K value and reduction experienced. The ability to separate classes has the highest Sum of FDR value obtained by K = 40% with a value of 6.94, higher than without reduction of 6.89.

3.2 Analysis

The graph of the results of the fifth test of the data set is presented in Figure 2. In the figure, the authors give the "x" mark to the highest Sum of FDR value. Data reduction on K-SVNN with certain K options turned out to provide high FDR value support as well. In the data sets tested by the authors, four of them provide a high FDR value when K is 30% to 50%.

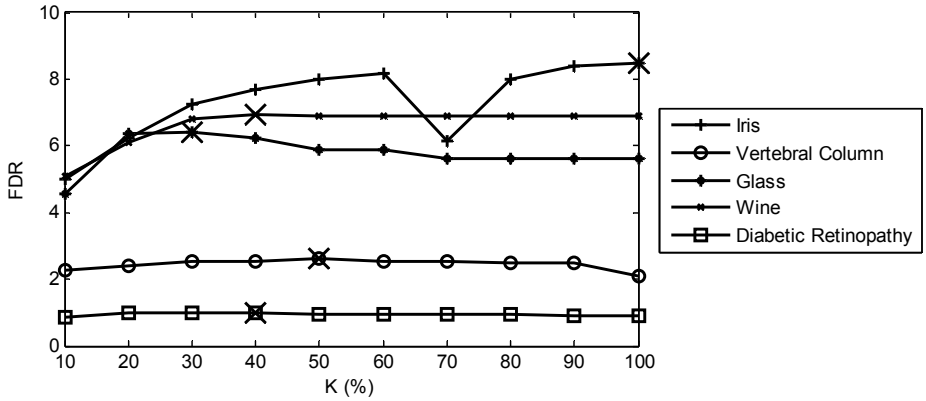

Figure 2. FDR value

At the highest FDR value, the reduction given to the K value is up to 4.49% as obtained in the Wine set data, while the smallest reduction is 0.26% as in the Diabetic Retinopathy data set. The graph of reduction rates obtained at the highest accuracy of the five data sets is presented in Figure 3.

For Iris data sets, the use of K-SVNN for data reduction does not increase the FDR value of all K options used. The value of FDR obtained is even lower than without reduction. Thus, for Iris data sets it is not appropriate to use K-SVNN for data reduction.

Table 4. Result of FDR from Wine data set

No.	K (%)	Reduction (%)	FDR of Feature													Sum of FDR
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	
1	10	49.44	1.39	0.01	0.29	0.02	0.20	0.03	0.00	0.01	0.00	2.05	0.13	0.01	0.93	5.09
2	20	20.22	1.73	0.15	0.25	0.08	0.29	0.00	0.01	0.00	0.00	2.14	0.23	0.06	1.17	6.10
3	30	11.24	2.05	0.22	0.27	0.06	0.36	0.01	0.00	0.00	0.00	2.26	0.27	0.07	1.25	6.81
4	40	4.49	2.23	0.23	0.24	0.07	0.26	0.00	0.01	0.00	0.01	2.24	0.32	0.09	1.25	**6.94**
5	50	1.69	2.30	0.20	0.28	0.07	0.18	0.00	0.00	0.00	0.01	2.23	0.29	0.09	1.22	6.88
6	60	1.12	2.32	0.20	0.28	0.07	0.18	0.00	0.00	0.00	0.01	2.25	0.27	0.09	1.23	6.91
7	70	0.00	2.33	0.20	0.29	0.07	0.18	0.00	0.00	0.00	0.01	2.18	0.30	0.09	1.23	6.89
8	80	0.00	2.33	0.20	0.29	0.07	0.18	0.00	0.00	0.00	0.01	2.18	0.30	0.09	1.23	6.89
9	90	0.00	2.33	0.20	0.29	0.07	0.18	0.00	0.00	0.00	0.01	2.18	0.30	0.09	1.23	6.89
10	100	0.00	2.33	0.20	0.29	0.07	0.18	0.00	0.00	0.00	0.01	2.18	0.30	0.09	1.23	6.89
11	Without reduction		2.33	0.20	0.29	0.07	0.18	0.00	0.00	0.00	0.01	2.18	0.30	0.09	1.23	6.89

Table 5. Result of FDR from Diabetic Retinopathy data set

No.	K (%)	Reduction (%)	FDR of Feature																	Sum of FDR
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	10	1.82	0.001	0.017	0.014	0.010	0.007	0.005	0.003	0.000	0.000	0.000	0.001	0.004	0.006	0.008	0.008	0.000	0.000	0.000	0.87
2	20	0.52	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.005	0.007	0.009	0.008	0.000	0.000	0.000	0.97
3	30	0.35	0.001	0.019	0.016	0.012	0.008	0.006	0.003	0.001	0.000	0.000	0.002	0.006	0.007	0.009	0.008	0.000	0.000	0.000	0.97
4	40	0.26	0.001	0.019	0.016	0.012	0.008	0.006	0.003	0.001	0.000	0.000	0.002	0.006	0.007	0.009	0.008	0.000	0.000	0.000	**0.98**
5	50	0.17	0.001	0.019	0.016	0.012	0.008	0.006	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.009	0.008	0.000	0.000	0.000	0.95
6	60	0.09	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.94
7	70	0.09	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.94
8	80	0.09	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.94
9	90	0.00	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.92
10	100	0.00	0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.92
11	Without reduction		0.001	0.019	0.016	0.012	0.008	0.005	0.003	0.001	0.000	0.000	0.002	0.004	0.005	0.008	0.007	0.000	0.000	0.000	0.92

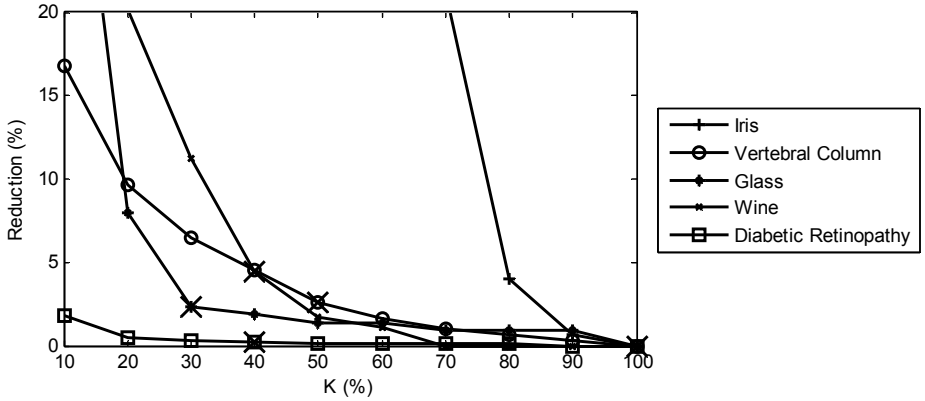

Figure 3. Reduction

Overall, the use of K-SVNN for data reduction does not necessarily increase the FDR value as a data capability parameter in separating classes. The FDR value here is still common among the classification methods. Need further testing on the classification method in order to obtain predictive accuracy.

4. CONCLUSION

The conclusions can be given from the results of this study as follows:

1. K-SVNN can be applied to the data set for data reduction while increasing the FDR value as a measure of the ability to separate classes.
2. In the five data sets used by the author, four of them can be applied K-SVNN for data reduction and one data set cannot. This means that in the data set used for classification it is necessary to test a number of K values of K-SVNN in order to know the effect of K-SVNN against it.

The suggestions are given as follows:

1. It is necessary to test the optimal choice of K on the classification method in order to know the performance increasing of the reduction performed.
2. The results of this study are limited to five data sets, need to be proved with other data sets to obtain a higher quality performance.

REFERENCES

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

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

[3] Prasetyo, E., Alim, S., and Rosyid, H. (2014). *Uji Kinerja dan Analisis K-Support Vector Nearest Neighbor Dengan SVM Dan ANN Back-Propagation*. in *Seminar Nasional Teknologi Informasi dan Aplikasinya*. Malang.

[4] Prasetyo, E. (2016). *K-Support Vector Nearest Neighbor: Classification Method, Data Reduction, and Performance Comparison.* Journal of Electrical Engineering and Computer Sciences. **1**(1): p. 1-6.

[5] Prasetyo, E. (2015). *Reduksi Data Latih Dengan K-SVNN Sebagai Pemrosesan Awal Pada ANN Back-Propagation Untuk Pengurangan Waktu Pelatihan.* SIMETRIS. **6**(2): p. 223-230.

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

[7] Theodoridis, S., and Koutroumbas, K. (2009). *Pattern Recognition*. Burlington,MA, USA: Academic Press.