






Randomspace-Based Fuzzy C-Means for Topic Detection on Indonesia Online News

Muhammad Rifky Yusdiansyah^(✉), Hendri Murfi,
and Arie Wibowo

Department of Mathematics, Universitas Indonesia, Depok 16424, Indonesia
{muhammad.rifky51, hendri, arie.wibowo}@sci.ui.ac.id

Abstract. Topic detection is a process used to analyze words in a collection of textual data to determine the topics in the collection, how they relate to each other, and how they change from time to time. Fuzzy C-Means (FCM) and Kernel-based Fuzzy C-Means (KFCM) method are clustering method that is often used in topic detection problems. Both FCM and KFCM can group dataset into multiple clusters on a low-dimensional dataset, but fail on high-dimensional dataset. To overcome this problem, dimension reduction is carried out on the dataset before topic detection is carried out using the FCM or KFCM method. In this study, the national news account's tweets dataset on Twitter were used for topic detection using the Randomspace-based Fuzzy C-Means (RFCM) method and Kernelized Randomspace-based Fuzzy C-Means (KRFCM) method. The RFCM and KRFCM learning methods are divided into two steps, which are reducing the dimension of the dataset into a lower-dimensional dataset using random projection and conducting the FCM learning method on the RFCM and the KFCM learning method on KRFCM. After obtaining the topics, then an evaluation is carried out by calculating the coherence value on the topics. The coherence value used in this study uses the Pointwise Mutual Information (PMI) unit. The study was conducted by comparing the average PMI values of RFCM and KRFCM with Eigenspace-based Fuzzy C-Means (EFCM) and Kernelized Eigenspace-based Fuzzy C-Means (KRFCM). The results obtained using national news account's tweets showed that the RFCM and KRFCM methods offered faster running time for a dimensional reduction but had smaller average PMI values compared to the average PMI values generated by the EFCM and KEFCM learning methods.

Keywords: Topic detection · Random projection · Fuzzy C-Means · Twitter

1 Introduction

Information and communication technology is overgrowing; this is characterized by the ease of obtaining information through the Internet. According to the data from the Internet Live Stats site, internet users in the world currently reach four billion, and the number continues to grow over time [1]. The rapid development of Internet technology is also followed by increasing the amount of information. Information can be obtained through social media, one of which is Twitter. Twitter allows the users to send messages, known as a tweet, about real-world events almost in real-time. More than

29 billion tweets are on Twitter and growing up to 300 million every day [2]. It makes Twitter becomes real-time information resource about real-world events locally in a region or globally in the world [3].

One of the information available on Twitter is a topic that is currently being discussed. To gather that information, we need some process called topic detection. Topic detection analyzes words in a collection of documents so that the topics contained in the document collection are found, how the topics relate to each other, and how the topic changes over time [4]. Topic detection can be done manually by reading all the textual data. However, due to a massive number of tweets, this manual way is difficult or even impossible. Therefore, a machine learning approach is needed to detect the topics.

One of topic detection methods is a clustering-based method. In the clustering method, the textual data is grouped in a way that members of the same cluster have more similarity to each other than with other group members. The centers of the clusters, called centroids, are means of their members. In topic detection problem, the centroids are interpreted as topics. K-means is a standard clustering method for topic detection [5–7]. This method splits the textual data into k clusters in which each textual data belongs to the nearest cluster center. In other words, the k -means method assumes that each textual data only belongs to one topic. This assumption is rather weak because most of the textual data have more than one topic. Therefore, soft clustering approach to be considered where each textual data may belong to more than one clusters. One of the soft clustering methods is fuzzy c-means (FCM) [8]. There is a modification of FCM using kernel trick called kernel-based fuzzy c-means (KFCM) [9]. Both FCM and KFCM works well on low dimensional data but fails on high dimensional data [10]. To overcome this problem, a method called Eigenspace-based Fuzzy C-Means (EFCM) and Kernelized Eigenspace-based Fuzzy C-Means (KEFCM) are proposed by transforming the high dimensional data into a lower-dimensional space using singular value decomposition and both methods are performed in this lower-dimensional space [11–13].

In this paper, we examine another projection method called random projection to transform the high dimensional data into lower-dimensional space. We called this method as randomspace-based fuzzy c-means (RFCM) and kernelized randomspace-based fuzzy c-means (KRFCM). We examine the performance of those methods for topic detection on Indonesian online news from Twitter. Our simulation shows that both RFCM and KRFCM achieve lower performance than both EFCM and KEFCM regarding topic interpretability in term of topic coherence. However, both RFCM and KRFCM are much more efficient and scalable.

The rest of the paper is organized as follows: Sect. 2 we present the reviews FCM, KFCM, GRP, RFCM, and KRFCM. Section 3 describes the results of our simulations and discussions. Finally, a general conclusion about the results is presented in Sect. 4.

2 Methods

Given textual data, the processes of clustering-based topic detection start by using the fuzzy c-means algorithm (FCM) [8] (Sect. 2.1) or kernel-based fuzzy c-means algorithm (KFCM) [9] (Sect. 2.2). Since textual data usually have high dimension and both FCM and KFCM fail in clustering the high dimensional data [10] Therefore, we use random projection (Sect. 2.3) to transform the textual data into the lower-dimensional space. We called this clustering-based topic detection algorithm as randspace-based fuzzy c-means (RFCM) (Sect. 2.4) and kernelized randspace-based fuzzy c-means (KRFCM) (Sect. 2.5).

2.1 Fuzzy C-Means

Fuzzy C-Means (FCM) produce some clusters by divide dataset into membership values $R = [r_{nk}]$, with r_{nk} values are between 0 and 1. The r_{nk} values implicitly say that one data can belong to one or more clusters. This method based on the minimization of the objective function as expressed:

$$J_{FCM}(R, M) = \sum_{k=1}^K \sum_{n=1}^N r_{nk}^w \|x_n - \mu_k\|^2 \quad (1)$$

where $\|\cdot\|$ is any norm represent similarities between the centroid and data, and w is a fuzzification constant to decide the level of fuzziness.

As the objective function is minimized, values in high membership area are given to data near the cluster centroid. The membership values and the centroids are computed iteratively using alternating optimization. Algorithm 1 below is the algorithm of fuzzy c-means.

Algorithm 1. Fuzzy C-Means [8]

Input : X , w , max number of iterations (T), threshold (ε)

Output : r_{nk}, μ_k

1. set $t = 0$
 2. initialize μ_k
 3. update $t = t + 1$
 4. calculate $r_{nk} = \left[\sum_{j=1}^K \left(\frac{\|x_n - v_k\|_2}{\|x_n - v_j\|_2} \right)^{2/w-1} \right]^{-1}, \forall n, k$
 5. calculate $\mu_k = \frac{\sum_{n=1}^N ((r_{nk})^w x_n)}{\sum_{n=1}^N (r_{nk})^w}, \forall k$
 6. if a stopping, i.e., $t > T$ or $\|R^t - R^{t-1}\|_F < \varepsilon$, is fulfilled then stop, else go back to step 3
-

2.2 Kernel-Based Fuzzy C-Means

Kernel-based Fuzzy C-Means (KFCM) works the same way as FCM but uses kernel function to measure the similarity between data and cluster centroid. The kernel function is used to make the clustering process performed in higher dimensional space without explicitly transforming the textual data.

Define nonlinear map as $\phi : x \rightarrow \phi(x) \in F$, where $x \in X$. X denotes the data space, and F the transformed feature space with a higher or even infinite dimension. Kernel function can be express as $ker(x, y) = \phi(x) \cdot \phi(y)$ which represents the inner product of ϕ function. Kernelized FCM minimizes the following objective function.

$$J_{KFCM}(R, M) = \sum_{k=1}^K \sum_{n=1}^N r_{nk}^w \|\phi(x_n) - \phi(\mu_k)\|^2 \tag{2}$$

where

$$\|\phi(x_n) - \phi(\mu_k)\|^2 = ker(x_n, x_n) + ker(\mu_k, \mu_k) - 2ker(x_n, \mu_k) \tag{3}$$

In simulations, we use RBF kernel functions, i.e., $ker(x, y) = exp(-\gamma\|x - y\|^2)$, so $ker(x, x) = 1$. According to Eq. 3, Eq. 2 can be written as

$$J_{KFCM}(R, M) = 2 \sum_{k=1}^K \sum_{n=1}^N r_{nk}^w (1 - ker(x_n, \mu_k)) \tag{4}$$

Minimizing Eq. 4 using alternating optimization under constraining M , we have

$$r_{nk} = \left(\frac{1}{\sum_{j=1}^K \left(\frac{1 - ker(x_n, \mu_k)}{1 - ker(x_n, \mu_j)} \right)} \right)^{\frac{1}{w-1}} \tag{5}$$

$$\mu_k = \frac{\sum_{n=1}^N r_{nk}^w ker(x_n, \mu_k) x_n}{\sum_{n=1}^N r_{nk}^w ker(x_n, \mu_k)} \tag{6}$$

As the objective function is minimized, values in high membership area are given to data near the cluster centroid. The membership values and the centroids are computed iteratively using alternating optimization. Algorithm 2 below is the algorithm of kernel-based fuzzy c-means.

Algorithm 2. Kernel-based Fuzzy C-Means [9]

 Input : X , w , max number of iterations (T), threshold (ε)

 Output : r_{nk}, μ_k

1. set $t = 0$
 2. initialize μ_k
 3. update $t = t + 1$
 4. calculate membership matrix using Eq. (5)
 5. calculate cluster centroid using Eq. (6)
 6. if a stopping, i.e., $t > T$ or $\|R^t - R^{t-1}\|_F < \varepsilon$, is fulfilled then stop, else go back to step 3
-

2.3 Random Projection

Random Projection is one of the reduction dimension methods for the matrix which has many uses in the data processing. There is no formal definition of Random Projection, but we will give the Definition 1 according to [14].

Definition 1. A random linear map $T : \mathbb{R}^m \rightarrow \mathbb{R}^k$ is called a random projection (or random matrix) if for all $\varepsilon \in (0, 1)$ and all vectors $x \in \mathbb{R}^m$, we have:

$$\mathbb{P}\left((1 - \varepsilon)\|x\|^2 \leq \|T(x)\|^2 \leq (1 + \varepsilon)\|x\|^2\right) \geq 1 - 2e^{-C\varepsilon^2 k} \quad (7)$$

For some universal constant $C > 0$ (independent of m, k, ε).

The random projection itself based upon the Johnson-Lindenstrauss lemma proposed in 1984 which state that “A set of points in a high-dimensional space can be projected into a lower-dimensional subspace of in such a way that relative distances between data points are nearly preserved” [15]. Note that the projected lower dimension subspace is selected randomly based on some distribution. In this case, we use Gaussian distribution. Algorithm 3 below is the algorithm of random projection.

Algorithm 3. Random Projection

 Input : X , data dimension (d), reduced dimension (c)

 Output: \tilde{X}

1. initialize random matrix $R_{c \times d}$ using $Gaussian(0, \frac{1}{c})$ distribution
 2. calculate \tilde{X} from multiplying $R_{c \times d}$ and X
 3. end
-

2.4 Randomspace-Based Fuzzy C-Means

Simply randomspace-based fuzzy c-means (RFCM) is to use random projection before using FCM. Random projection reduces the dimension of data X , as described in Algorithm 3. As mentioned above, the output of random projection is reduced matrix \tilde{X} .

Next, we perform FCM using matrix on the reduced dimensional data \tilde{X} . In this step, we calculate the membership matrix R based on \tilde{X} . To calculate the centroid matrix, we multiply the membership matrix R with the original data X . Algorithm 4 below is the algorithm of randomspace-based fuzzy c-means.

Algorithm 4. Randomspace-based Fuzzy C-Means

Input : X , w , data dimension (d), reduced dimension (c), max number of iterations (T), threshold (ε)

Output : r_{nk}, μ_k

1. set $t = 0$
 2. reduce the dimension of the data X to c -dimensional data \tilde{X} using random projection (Algorithm 3)
 3. initialize μ_k
 4. update $t = t + 1$
 5. calculate $r_{nk} = \left[\sum_{j=1}^K \left(\frac{\|x_n - v_k\|_2}{\|x_n - v_j\|_2} \right)^{2/w-1} \right]^{-1}, \forall n, k$
 6. calculate $\mu_k = \frac{\sum_{n=1}^N ((r_{nk})^w x_n)}{\sum_{n=1}^N (r_{nk})^w}, \forall k$
 7. if a stopping, i.e., $t > T$ or $\|R^t - R^{t-1}\|_F < \varepsilon$, is fulfilled then stop, else go back to step 4
 8. calculate cluster centroid from original data with multiplying R and original data X
-

2.5 Kernelized Randomspace-Based Fuzzy C-Means

Same as RFCM, kernelized randomspace-based fuzzy c-means (KRFCM) is to use random projection before using KFCM. Random projection reduces the dimension of data X , as described in Algorithm 3. As mentioned above, the output of random projection X is reduced matrix \tilde{X} .

Next, we perform FCM using matrix on the reduced dimensional data \tilde{X} . In this step, we calculate the membership matrix R based on \tilde{X} . To calculate the centroid matrix, we multiply the membership matrix R with the original data X .

Algorithm 5. Kernelized Randomspace-based Fuzzy C-Means

Input : X, w , data dimension (d), reduced dimension (c), max number of iterations (T), threshold (ε)

Output : r_{nk}, μ_k

1. set $t = 0$
 2. reduce the dimension of the data X to c -dimensional data \tilde{X} using random projection (Algorithm 3)
 3. initialize μ_k
 4. update $t = t + 1$
 5. calculate membership matrix using Eq. (5)
 6. calculate cluster centroid using Eq. (6)
 7. if a stopping, i.e., $t > T$ or $\|R^t - R^{t-1}\|_F < \varepsilon$, is fulfilled then stop, else go back to step 4
 8. calculate cluster centroid from original data with multiplying R and original data X
-

3 Results and Discussion

In this section, we are going to analyze and compare the accuracies of the RFCM and KRFCM with the Eigenspace-based Fuzzy C-Means (EFCM) [11] and Kernelized Eigenspace-based Fuzzy C-Means (KEFCM) [13] method. The comparison uses a measurement unit called *topic interpretability*.

3.1 Topic Interpretability

Topic Interpretability is a quantitative method to seek the interpretability of a topic by calculating the coherence scores of words that construct the topic. One of the standard formulas which used to estimate coherence scores is PMI. Suppose a topic t consists of an n -word that is $\{w_1, w_2, \dots, w_n\}$, the PMI of the topic t is

$$PMI(t) = \sum_{j=2}^n \sum_{i=1}^{j-1} \log \left(\frac{P(w_j, w_i)}{P(w_i)P(w_j)} \right) \quad (8)$$

where $P(w_j, w_i)$ is the probability of the word w_i appears together with the word w_j on the corpus, $P(w_i)$ is the probability of the word w_i appears in the corpus, and $P(w_j)$ is the probability of the word w_j appears in the corpus. Corpus is a database consisting of some text-based documents which used for a reference to calculate PMI. In this experiment, we use a corpus consisting of 3.2 million Indonesian Wikipedia documents.

In this simulation, we use the dataset from 14 Indonesian national news account on twitter with the majority are the Indonesian language. The dataset is gathered by using a streaming method from March 4th, 2019 until March 9th, 2019. The streaming method itself is done by taking the tweets after the tweets are being made. The benefit of the streaming method is that the data gathered are live tweets. The datasets consist of 5425 tweets and 11056 words. Table 1 below provides some additional information about the dataset used in this paper.

Table 1. Additional Information about the dataset

Number of tweets	5425
Number of words	11056
Number of national news account	14
Streaming start (DD/MM/YYYY HH:MM:SS)	04/03/2019 15:07:28 GMT + 07
Streaming end (DD/MM/YYYY HH:MM:SS)	09/03/2019 06:36:14 GMT + 07
Streaming duration (hours)	111.5

The datasets have been in the form of a list of sentences. So, the process that needs to be done is only to form the word-document matrix X and do the weighting process. In this simulation, the weighting process is performed using the term frequency-inverse document frequency (TF-IDF) [16].

To convert text into vector representations, we do some pre-processing. Firstly, we delete link on every tweet; convert some words with the hashtag to the separate words; convert all words into lowercase; replace two or more repeating letters with only two occurrences; delete all the numbers on tweets; delete every username on tweets; delete the words with two or fewer characters. For the final step, we used the term frequency-inverse document frequency for weighting the dataset. Besides using scikit-learn packages for preprocessing and tokenizing steps, we also used the natural language toolkit from nltk [17].

In this simulation, we also inspect two other methods for topic detection; they are EFCM [11] and KEFCM [13]. For EFCM, KEFCM, RFCM, and KRFCM parameters, we arrange the fuzzification constant $w = 1.5$, the maximum number of iteration $T = 1000$, the threshold $\varepsilon = 0.005$ and we are reducing its dimension into 5 using Truncated SVD for EFCM and KEFCM and Gaussian Random Projection for RFCM and KRFCM. As for KEFCM and KRFCM, we use $\gamma = 0.001$. All of these methods are repeated five times.

These methods produce topics consisting of 10 words, in each topic forming one vector. And after the results are collected, then the words in each topic are going to be calculated by PMI in Eq. 8 to get coherence score.

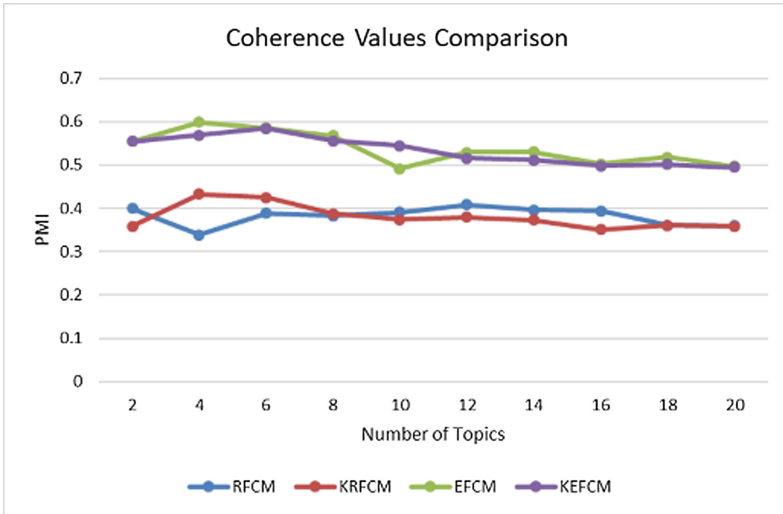


Fig. 1. Comparison in PMI score between RFCM, KRFCM, EFCM, and KEFCM

Figure 1 shows the comparison of Coherence scores for the number of topic $c \in \{2, 4, \dots, 18, 20\}$. Based on Fig. 1, we can see that RFCM and EFCM have smaller PMI values than EFCM and KEFCM methods. Table 2 below gives a summary of the PMI scores for each method. EFCM method achieves the highest optimal PMI scores of the other three methods.

Table 2. The comparison of the optimal topic interpretability (PMI) scores

RFCM	KRFCM	EFCM	KEFCM
0.4087	0.4331	0.5471	0.5468

Figure 2 below shows that the random projection outperformed the SVD in terms of running time.

Analytically, the running time of the random projection can be denoted as $O(c \times k \times N)$ where c states the number of entities not empty per column, k denotes the dimensions after being reduced, and N denotes the amount of data [18]. While the running time of the SVD can be denoted as $O(n_{max}^2 \times k)$ with $n_{max} = \max(n_{document}, n_{feature})$ where the $n_{document}$ states the number of documents and $n_{feature}$ state the number of different words and k denotes the dimensions after being reduced¹.

¹ <https://scikit-learn.org/stable/modules/decomposition.html>.

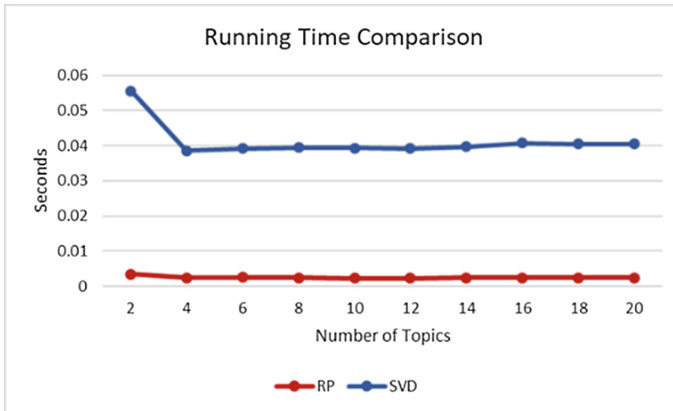


Fig. 2. Comparison in running time between random projection and SVD

4 Conclusions

In this paper, we examine randomspace-based fuzzy c-means and kernelized randomspace-based fuzzy c-means for topic detection on twitter. We use interpretability of the extracted topics in the form of coherence scores to compare RFCM and KRFCM methods with EFCM and KEFCM methods. Our simulations show that RFCM gives best results for 12 topics, and KRFCM gives best results for 4 topics. However, Both the RFCM and KRFCM learning methods cannot outperform the average PMI values produced by the EFCM or KEFCM learning methods, but RFCM and KRFCM excel at the time needed to reduce the dimensions of the dataset.

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

References

1. Xie, W., Zhu, F., Jiang, J., Lim, E.-P., Wang, K.: Topic sketch: real-time bursty topic detection from Twitter. *IEEE Trans. Knowl. Data Eng.* **28**(8), 2216–2229 (2016)
2. Craig, T., Ludloff, E.M.: *Privacy and Big Data*. O'Reilly Media Inc., Sebastopol (2011)
3. Aiello, L.M., et al.: Sensing trending topics in Twitter. *IEEE Trans. Multimedia* **15**(6), 1268–1282 (2013). <https://doi.org/10.1109/TMM.2013.2265080>
4. Blei, D.M.: Probabilistic topic models. *Commun. ACM* **55**(4), 77–84 (2012)
5. Petkos, G., Papadopoulos, S., Kompatsiaris, Y.: Two-level message clustering for topic detection in Twitter. In: *Proceedings of the SNOW 2014 Data Challenge*, Seoul, Korea, 8 April 2014 (2014)

6. Nur'aini, K., Najahaty, I., Hidayati, L., Murfi, H., Nurrohmah, S.: Combination of singular value decomposition and k-means clustering method for topic detection on Twitter. In: Proceedings of International Conference on Advanced Computer Science and Information System, Depok, Indonesia, 10–11 October 2015 (2015)
7. Fitriyani, S.R., Murfi, H.: The k-means with mini batch algorithm for topics detection on online news. In: Proceedings of the 4th International Conference on Information and Communication Technology, Bandung, Indonesia, 25–27 May 2016 (2016)
8. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Platinum Press, New York (1981)
9. Daniel, G., Witold, P.: Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study. *Fuzzy Sets Syst.* **161**(3), 522–543 (2010). <https://doi.org/10.1016/j.fss.2009.10.021>
10. Winkler, R., Klawonn, F., Kruse, R.: Fuzzy c means in high dimensional spaces. *Int. J. Fuzzy Syst. Appl.* **1**, 1–16 (2011)
11. Muliawati, T., Murfi, H.: Eigenspace-based fuzzy c-means for sensing trending topics in Twitter. In: AIP Conference Proceedings, vol. 1862, no. 1, July 2017. <http://doi.org/10.1063/1.4991244>
12. Murfi, H.: The accuracy of fuzzy c-means in lower-dimensional space for topic detection. In: Qiu, M. (ed.) *SmartCom 2018*. LNCS, vol. 11344, pp. 321–334. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-05755-8_32
13. Prakoso, Y., Murfi, H., Wibowo, A.: Kernelized eigenspace based fuzzy C means for sensing trending topics on Twitter. In: Proceedings of the International Conference on Data Science and Information Technology, Singapore (2018)
14. Vu, K.K.: Random projection for high-dimensional optimization. *Optimization and Control*. Université Paris-Saclay. English (2016)
15. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mapping into Hilbert space. In: Conference in Modern Analysis and Probability. *Contemporary Mathematics*, vol. 26, pp. 189–206. American Mathematical Society (1984)
16. Manning, C.D., Schuetze, H., Raghavan, P.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
17. Loper, E., Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL Interact. Present. Sess, pp. 69–72 (2006)
18. Bingham, E., Mannila, H.: Random projection in dimensionality reduction. In: Proceeding of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York (2001)