Conference Paper

Determining and Clustering Potential Legislative Candidate in West Java District using K-Nearest Neighbors Algorithm

Faiza Renaldi, Alfin Dhuhawan Bagja, and Gunawan Abdillah

Department of Computer Science, Faculty of Mathematics and Natural Science, Jenderal Achmad Yani University, Cimahi – Jawa Barat, Indonesia

Abstract

Indonesia held its first general election in 1955 to elect legislatures from all provinces. The latest was held in 2014, which elected 560 members to the People's Representative Council (Dewan Perwakilan Rakyat, DPR) and 128 to the Regional Representative Council (Dewan Perwakilan Daerah, DPD). The PRC was elected by proportional representation from multi-candidate constituencies/districts. Currently, there are 77 constituencies in Indonesia, each of which returns 3-10 Members of Parliament based on population. Under Indonesia's new multi-party system, no party has been able to secure an outright victory; hence, selecting the right candidate for the right constituencies has been a major effort for all participating parties. Many combinations have been tried; popularities, intelligence, public figures, 'putera daerah' are all variables that can only show a fraction of winning pattern where no general conclusion can be drawn. This research used data mining techniques to create an unfound pattern, and to suggest which particular legislative candidate is most suitable for which constituency. Using 11 West Java constituencies (11 clusters), K-Nearest Neighbors (K-NN) algorithms, we found out that an 83.33% accuracy using data from 2014 general election.

Keywords: Constituencies, data mining, election strategy, KNN algorithm

Corresponding Author: Faiza Renaldi; email: faiza.renaldi@unjani.ac.id

Received: 09 April 2017
Accepted: 17 May 2017
Published: 12 June 2017

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICoSaPS Conference Committee.

OPEN ACCESS

1. Background

There was a time when popularity was considered an important asset to become a legislator in DPR, hence many artistes chose to leave their career for politics. That is not the case now. Some well-known names such as Sandy Nayoan (actor) and former world badminton champion, Ricky Subagja, were not able to become a legislator. To make things worse, even artistes who were qualified for parliament in the previous period were eliminated from the competition (Ingrid Kansil and Nurul Arifin). Well-known politicians such as Marzuki Alie, Priyo Budi Santoso, Eva Kusuma Sundari, and Sutan Bhatoegana also failed in the elections. This proves that the constituent's preferences are getting more and more diverse, and parties should build more scientific

and appropriate strategies to achieve their competitive advantage to compete in the general election.

One of the key strategies is to determine which regions suits one particular candidate by data mining the pattern of previous legislator's personal information. Data mining is defined as a process of pattern discovery in a set of data [7] known to be able to predict an outcome of a certain dataset. Previous studies of data mining methods have been used to determine motorcycle ownership credit risk [7] and credit application approval [12], to classify job opportunities of fresh graduates [11], to determine loan approval for cooperation [9], to predict university graduates [1], and to predict loan customer category [8].

In this research, we use KNN algorithm to match reference data that were stored in designated clusters with a new data to predict its outcome. This algorithm has also been widely used in previous studies including data classification results of palm oil production of PT. Minamas Kecamatan Parindu [5], hotspot classification on peat lands in Sumatera and Kalimantan [6], determination of scholarship's recipients [13], identification of batik's pattern [15], news text classification [4], and even on cardiac disease diagnosis [14]. Data used as reference are national house of representative (DPR) legislative member of 2009-2014, 2014-2019. Regions or Districts are specified to 11 constituencies of West Java.

This paper aims to determine whether or not such algorithm can be used to determine the possibility of one particular candidate to be clustered into one particular district or constituency. In the context of this study, the research questions are:

RQ1: How effective is the KNN-Algorithmbeing used to cluster and predict legislative candidates for their respective constituencies especially in West Java District?

RQ2: Does the algorithm accurately predict the election of particular legislators compared with previous general election data (2014)?

We expect that the use of algorithm and eventually computer based information systems of candidate registration and selection can be tested and implemented soon to political parties in order to enhance their selection mechanism of legislative candidates, and eventually increase their level of success to enter into the House of Representative.

2. Literature Review

2.1. Data Mining

Data mining is used to uncover key information in the database. Data mining is also defined as a process of discovering trends in a dataset, or as a series of processes for adding value in the form of knowledge that were unknown from a dataset [3]. Data

mining is a part of the knowledge discovery process in databases (KDD) in charge of extracting patterns or models of data by using a specific algorithm. The process of KDD is as follows:

1. Data selection: selection of data from the operational dataset needs to be done before the stage of extracting information in KDD.
2. Cleansing: removing data duplication, check data for inconsistencies, and correct errors in data such as printing errors.
3. Transformation: transform data into a one generic format for the data mining process.
4. Data mining: search for patterns or interesting information in the selected data by using techniques or methods (in this case, we will use K-NN Algorithm).
5. Evaluation: The information pattern is derived from data mining process needs to be presented in a form easily understood by stakeholders.

2.2. K-Nearest Neighbors

K-NN algorithms belong to the supervised algorithms obtained through a learning process (learning) upon reference data that has been classified, and learning outcomes are used to classify new data with unknown output. In the K-NN algorithm, a new data is classified based on the distance of the new data with new data similarity level closest to the data pattern [10]. There are many similarities in the K-Nearest Neighbors algorithm to determine the distance between the proximity of old data with new data, one of which is the Euclidean distance. It is formulated in the following equation:

$$d_i = \sqrt{\sum_{i=1}^{14} (x_{2i} - x_{1i})} \quad (1)$$

Where:

x_1 = reference data

x_2 = test data

i = Individual attributes between 1 to 14

d = distance

3. Method - Data Collection

We gathered the data of legislators who entered the parliament in 2004 and 2009. The data were extracted from online sources such as political parties' websites, Wikipedia,

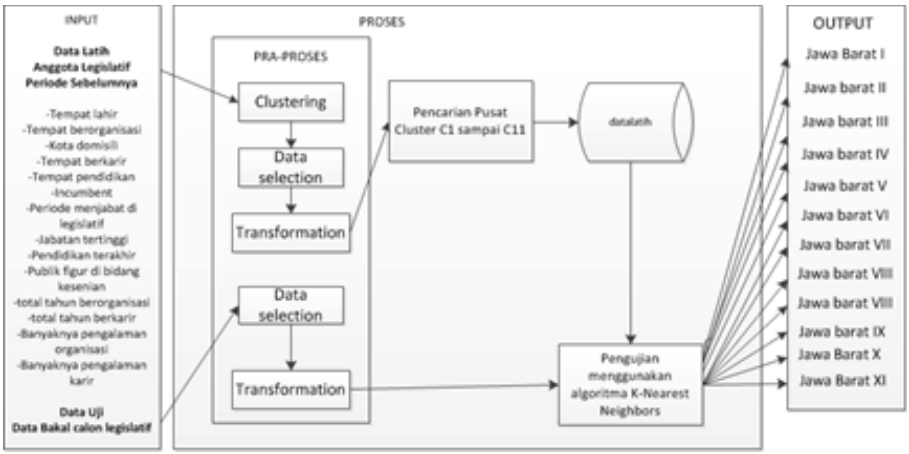

Figure 1

and also reference books on DPR of Indonesia. There were 91 people in 2009 and 91 people in 2014 who became legislators leading to 182 data references. However, since there were incumbents in 2009 and some data were not showing detailed information and therefore could not be treated as references, we decided to only use 142 of them.

4. Requirement Analysis

This stage involved interviews with two political parties (NASDEM and GOLKAR of West Java) to derive information on how they chose their candidates for the legislative assembly. Through a series of interviews, we concluded that there were 3 main processes: registration – selection – classification. Firstly, the candidates registered themselves at the office. Then, a series of events was conducted to select which candidate suited one party's criteria. Finally, the candidate would be classified according to the list of constituencies available in West Java.

4.1. System Design

The suggested system is depicted in Figure 1.

Before using the equation 1, we first clustered data reference based on 11 constituencies/regions of elections in West Java. After the group was formed, the next step was to find the center of the clusters, which then were counted the proximity to equation 1.

4.2. Software Design and Implementation

In this stage, we conducted the design and then implemented the calculation in a software system and produced an output that determined the successful prediction,

Figure 2

and then chose whichregion/constituency one particular candidate was most likely to win in the legislative elections of West Java.

4.3. Testing and Evaluation

We tested the accuracy of the algorithm by entering data from previous election to see whether or not the result was the same as on the ground. The findings would be presented in the conclusion section of this paper.

5. Findings and Discussion

We set 14 inputs to the system: (1) place of birth, (2) region of association, (3) place of education, (4) domicile city, (5) career city, (6) incumbent (legislators are still in

Clusters	Distance Result
C1	$\sqrt{(19-20.25)^2+(19-20.83)^2+(28-23)^2+(28-23.17)^2+(19-21.17)^2+(1-1.25)^2+(0-3.08)^2+(7-4.92)^2+(3-3.5)^2+(1-1.17)^2+(5-9.08)^2+(6-7.75)^2+(8-4.58)^2+(3-4.25)^2}$=**10.25**
C2	$\sqrt{(19-21.87)^2+(19-4.27)^2+(28-21.47)^2+(28-19.8)^2+(19-13.73)^2+(1-1.27)^2+(0-6.07)^2+(7-5.73)^2+(3-3.4)^2+(1-1.27)^2+(5-9.6)^2+(6-8.53)^2+(8-5.8)^2+(3-3.93)^2}$=**20.85**
C3	$\sqrt{(19-21)^2+(19-18.63)^2+(28-24.13)^2+(28-16.31)^2+(19-22)^2+(1-1.44)^2+(0-4.69)^2+(7-4.13)^2+(3-3.69)^2+(1-1)^2+(5-7.94)^2+(6-8.31)^2+(8-4.38)^2+(3-4.75)^2}$=**15.03**
C4	$\sqrt{(19-24.38)^2+(19-19)^2+(28-26.13)^2+(28-25.13)^2+(19-22.38)^2+(1-1.25)^2+(0-2.88)^2+(7-5.25)^2+(3-3.88)^2+(1-1.25)^2+(5-9)^2+(6-9.63)^2+(8-5.75)^2+(3-6.125)^2}$=**10.41**
C5	$\sqrt{(19-23.86)^2+(19-8.93)^2+(28-24.43)^2+(28-15.43)^2+(19-15.5)^2+(1-1.43)^2+(0-5.07)^2+(7-4.36)^2+(3-3.57)^2+(1-1.14)^2+(5-9)^2+(6-9.36)^2+(8-5.5)^2+(3-5.07)^2}$=**19.47**
C6	$\sqrt{(19-28.22)^2+(19-23.56)^2+(28-27.89)^2+(28-24.33)^2+(19-26)^2+(1-1.11)^2+(0-1.78)^2+(7-3.22)^2+(3-3.11)^2+(1-1.11)^2+(5-7.44)^2+(6-9.78)^2+(8-4.56)^2+(3-5.67)^2}$=**14.5**
C7	$\sqrt{(19-22.33)^2+(19-8.2)^2+(28-25.33)^2+(28-15.27)^2+(19-17.93)^2+(1-1.13)^2+(0-2)^2+(7-5.6)^2+(3-3.67)^2+(1-1.07)^2+(5-7.27)^2+(6-7.87)^2+(8-4.4)^2+(3-3.47)^2}$=**18.06**
C8	$\sqrt{(19-19.08)^2+(19-14.77)^2+(28-23.62)^2+(28-21.62)^2+(19-15.77)^2+(1-1.38)^2+(0-6.38)^2+(7-4.23)^2+(3-3.62)^2+(1-1.15)^2+(5-8.85)^2+(6-7.23)^2+(8-5.23)^2+(3-4.23)^2}$=**12.75**
C9	$\sqrt{(19-18.45)^2+(19-16.45)^2+(28-17.45)^2+(28-21.27)^2+(19-15.55)^2+(1-1.27)^2+(0-3.27)^2+(7-5.64)^2+(3-3.18)^2+(1-1.09)^2+(5-8.45)^2+(6-8.82)^2+(8-6.27)^2+(3-4.18)^2}$=**14.57**
C10	$\sqrt{(19-17.73)^2+(19-10.55)^2+(28-25.18)^2+(28-19.27)^2+(19-16.27)^2+(1-1.18)^2+(0-3.64)^2+(7-3.09)^2+(3-3.36)^2+(1-1)^2+(5-7.36)^2+(6-8.27)^2+(8-4.27)^2+(3-4.91)^2}$=**14.89**
C11	$\sqrt{(19-23.22)^2+(19-14.5)^2+(28-23.72)^2+(28-23.39)^2+(19-19.83)^2+(1-1.28)^2+(0-3.83)^2+(7-6.11)^2+(3-3.5)^2+(1-1)^2+(5-9.83)^2+(6-9.22)^2+(8-5.78)^2+(3-5.44)^2}$=**11.78**

Table 1

office and then again ran in the general election), (7) total year as a legislature, (8) highest rank in the office, (9) highest/latest education, (10) public figures in the arts, (11) total years of organizational experience, (12) total career, (13) number of previous organization, and (14) number of career.

To visualize the input process, we tested it with one dataset under the name of Handoko Suriaatmaja, in accordance to the master fields above: (1) Kota Bandung,

(2) Kota Bandung, (3) Jakarta, (4) Jakarta, (5) Kota Bandung, (6) No, (7) None, (8) Board/Chairman, (9) Undergraduate, (10) No, (11) 4, (12) 5, (13) 7, (14) 2.

Next, we compared the test data with pre-defined clusters, which are set in accordance to the 11 constituencies of West Java. Using K-Nearest Neighbors algorithm, we then calculated the Euclidian distance of our test data compare to the 11 clusters, which was pre-defined earlier. The output of this system in the form of recommendations for potential electoral candidates was based on the value of the smallest distance to the center of clusters. The calculation can be shown in Table 1:

The smallest distance found in Clusters 1 with a value of 10.25, making recommendations for the electoral district Handoko Suriaatmaja is constituency of **West Java I**. This action can be performed repeatedly until the total slots on each regions/constituency are filled in accordingly.

The implementation of this research has resulted in software which can be seen in a set of user interfaces in Figure 2.

Our research has proven to be able to give output and recommendations, and even shown the relative accuracy along the way. However, there is still a much work in to be done, in which some limitations occur such as the absence of interactive communication between candidates and parties, while all data entries are being handled by an administrator.

6. Conclusion

This research has resulted in a system that can determine the success of potential legislative candidates in the electoral district of West Java using K-NN Algorithm. We also performed accuracy test on 12 different previous legislators, and resulted in 10 correct and 2 incorrect results; hence we determine that the accuracy of this system is 83.33%.

We believe our analysis can bring much useful information to political parties participating in the general election and also to the future candidates for legislative office. However, our study is still a work in progress in many areas; hence further tests and developments need to be carried out. One suggestion is to add not only the successful legislators but also the failed ones so it can bring more data variances in each constituency.

References

[1] Yesi Andri, N. K, and M. Sri, Seminar Nasional Informatika, ISSN: 1979-2328, Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa

Pada Universitas Bina Darma Palembang. Seminar Nasional Informatika ISS, 1979-2328, 2013.

[2] M. Ayub, K. Tanti, and C. Maresha, "Model Analisis Classification Dengan," in J48 Untuk Data Mahasiswa dan Dosen Di Perguruan Tinggi. ISSN: 1979-3960, pp. 19–30, SNASTIA, 1979-3960, 2014.

[3] S. L. B. Ginting, W. Zarman, and A. Darmawan, "Teknik Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighborhood," Jurnal Teknik Komputer Unikom, vol. 3, no. 2, pp. 29–34, 2014.

[4] Edi Junadi, 2015. Penggunaan KNN (K-Nearest Neighbor) Untuk Klasifikasi Teks Berita Yang Tak-Terkelompokkan Pada Saat Pengklasteran Oleh STC (Suffix Tree Clustering). Vol. 9, ISSN: 1979-8911.

[5] N. Krisandi, Helmi, and B. Prihandono, "Algoritma K-Nearest Neighbors Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit pada PT. Minamas Kecamatan Parindu," Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster), vol. 2, no. 1, pp. 33–38, 2013.

[6] F. Kusumaningrum and S. S. Imas, "Klasifikasi Kemunculan Titik Panas pada Lahan Gambut di Sumatera dan Kalimantan Menggunakan Algoritma K-Nearest Neighbor," Makalah Koloklum Program S1 Ilmu Komputer Alih Jenis, 2013.

[7] H. Leidyana, "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor. Jurnal Penelitian Ilmu Komputer," System embedded & Logic, vol. 1, no. 1, pp. 65–76, 2013.

[8] A. G. Mabrur and L. Riani, "Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit," Jurnal Komputerdan Informatika (KOMPUTA), 2012.

[9] I. Menarianti, "Klasifikasi Data Mining Dalam Menentukan Pemberian Kredit Bagi Nasabah Koperasi," Jurnal Ilmiah teknosains, vol. 1, 2015.

[10] R. I. Ndaumanu, Kusrini, and R. M. Arief, "Kusrini Arief," Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor. Jatisi, vol. 1, no. 1, pp. 1–15, 2014.

[11] Nursalim. Suprapedi and H. Himawan, "Klasifikasi Bidang Kerja Lulusan Menggunakan Algoritma K-Nearest Neighbor," Jurnal Tekonologi Informasi, vol. 10, no. 1, pp. 31–43, 2014.

[12] E. S. Y. Pandie, "Implementasi Algoritma Data Mining K-Nearest Neighbour (K-NN) Dalam Pengambilan Keputusan Pengajuan Kredit," Seminar Nasional Sains dan teknik, pp. 31–34, 2012.

[13] H. Risman, N. Didik, and R. W. Yustina, "Penerapan Metode K-Nearest Neighbor Pada Aplikasi Penentu Penerima Beasiswa Mahasiswa Di STMIK Sinar Nusantara Surakarta," ISSN: 2338-4018, pp. 2338–4018, 2013.

[14] M. Shouman, T. Tim, and S. Rob, "Applying K-Nearest Neighbor in Diagnosing Heart Disease Patients," in *International Journal of Information and Education Technology 2*, vol. 2, p. 3, 2012.

[15] J. W. Yodha and W. K. Achmad, "Pengenalan Motif Batik Menggunakan Deteksi Tepi Canny dan K-Nearest Neighbor," Techno.com, vol. 13, pp. 251–262, 2014.