Clustering of Social Network Data by Means of Kohonen Neural Networks

Galim Vakhitov [a], Zulfira Enikeeva [a], Pavel Ustin[a], Nagim Davletshin [a]

[a] *Kazan Federal University, 18 Kremlevskaya street, Kazan, 420111, Russian Federation*

Abstract

Social network user profile analysis is becoming an increasingly common source of information. In particular, the profile analysis can be used when hiring young employees. This information is poorly structured and it cannot be statistically analyzed in terms of what social network profile corresponds to a promising employee who has previously been a successful student.

In this regard, we have carried out the clustering of these student profiles in the social network using the Kohonen neural network and revealed connections of the clusters with the data concerning the academic performance of students included in these clusters.

Keywords [1]

Social networks, academic performance, data clustering, the Kohonen neural network

1. Introduction

Various machine learning techniques are used for data clustering. Neural networks are one of the most popular of them. Their use is well-grounded, especially if the data is poorly structured. Social networks users' profile data is an example of such data. The attempts to cluster social networks users based on information of qualitative nature have already been made. In particular, there are works on the clustering of users based on the communities to which they belong [1], [2]. However, the great part of the information on a social network user's profile is quantitative. For example, not only the contents of photos or videos on the user's page can be examined, but the number of these photos and videos and their proportion as well. The list of friends and subscribers can be analyzed, as well as the number of friends and subscribers and their proportion, and conclusions about the user himself can be obtained based on this quantitative information.

The goal was to substantiate our position that on the page of social network users it is possible to identify metrics with quantitative data, the values of which and the ratio between which may contain information about the user of the social network.

This article describes the results of our research work on the use of the Kohonen neural network [3], [4] for identification of clusters among social network users who were university students. For this purpose the quantitative information of the profiles was analyzed.

2. Methods

We carried out the research study of the relationship between the quantitative metrics of the personal profile of students (in the sample group of 33484 people) in the social network Vkontakte and their academic performance.

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia
EMAIL: GZVahitov@kpfu,ru (Galim Vakhitov); ZAEnikeeva@kpfu.ru (Zulfira Enikeeva); Pavel.Ustin@kpfu.ru (Pavel Ustin); davlet-9@yandex.ru (Nagim Davletshin)
ORCID: 0000-0002-8953-728X (Galim Vakhitov); 0000-0002-2468-7455 (Zulfira Enikeeva); 0000-0003-3950-743 (Pavel Ustin); 0000-0002-5807-8527 (Nagim Davletshin)

The numerical data in terms of the quantity of the elements (friends, subscribers, photos, videos, interesting pages) on the user's page was chosen as metrics. That is, every student - on the one hand - was characterized by one indicator representing his academic performance, and on the other hand, he was characterized by the set of 5 numerical parameters. All data of students was anonymous. The most informative numerical parameters were identified as a result of correlation analysis. In total, at least 20 numeric parameters can be obtained in this social network.

Academic performance is an individual characteristic of every student. According to this parameter the students were divided into three groups: students with good academic performance, students with satisfactory academic performance, students with bad academic performance. To distinguish these three categories of students, we carried out statistical processing of data on the marks received by students on exams and tests during their studies. The choice of certain estimates for the formation of the indicator of academic success was justified.

We analyzed the structure of students' groups, the assessment practice of teachers in order to establish the differences between students with good, satisfactory and bad academic performance.

We also highlighted the students for whom the social network is rather significant, and we can use the qualitative and quantitative data of their profile as data for our research work.

The clusterization of data obtained from the social network Vkontakte was carried out using Kohonen neural networks. It was clustered into 2, 3, 4 clusters. In some cases the highlighted clusters identified students with good and bad academic performance with high accuracy. So, we draw the conclusion the possibility to use Kohonen neural networks for clustering the poorly structured data of students' profiles in the social network Vkontakte .

Then we characterize the methodology of dividing the university students into two groups: with good academic performance and bad academic performance. This methodology is determined by the learning process at the university and it is likely that in the other universities it will be different.

We made a decision not to use directly the marks in points received by students in the test and exam, but to use these qualitative concepts of academically successful and academically unsuccessful students for a number of reasons. The main one is that students, being in groups, are not guided by the grades received for passing tests and exams. Students try to hold onto their position in group performance. Firstly, each student gets used to the fact that he is a better student and worse than any other student, and strives to maintain his place in the group. Secondly, if the information about the student's progress in the current semester did not allow us to assign him to a certain group in terms of his progress, then we turned to the data on his progress in the previous semester.

In the learning process the students are organized into academic groups. there are from 2 to 40 students in every group. For our research study we used the data of students enrolled in groups of 5 or more people, as far as in groups of 2-4 people it is impossible to ensure data anonymity and to highlight the best and worst students, since the grades differ insignificantly.

Students study 4 years (Bachelor's programmes), then 2 years — Master degree programmes. Also, some students study 5 years. KFU includes institutes and one faculty. Educational process is organized within semesters. Every academic year includes two semesters. Bachelors receive grades within the course of 8 semesters, Masters — 4 semesters, specialists - 10 semesters. In some fields and specialties the education is carried out in the external study mode as well. In this case the number of educational semesters increases.

The study of a number of disciplines finishes at the end of the semester by a test. More important — from the point of view of the university — disciplines are finished by the exams. The marks received by a student after passing the tests and exams during the sessions are the criterion of the academic performance. Students get grades during the semester, and then they additionally receive grades when taking tests or exams. Finally — based on the sum of these points — students get the finally mark: excellent (if the total number of grades is from 86 to 100), good (71-85), satisfactory (56-70), unsatisfactory (the number of grades is less than 56).

We used information of the total sum of grades received by a student when taking the exams in order to divide students into groups with good, satisfactory and bad academic performance.

This is due to the fact that students care more about the principal result — to pass the exam, they do not care about the number of grades.

Now we will briefly explain why it is impossible to single out successful and unsuccessful students, regardless of the areas of study and courses within the university.

As an assessment mark of the successful educational course we can use information about the grades received by a student in the academic disciplines only in connection with the information about the course of study, the student group, the fields of study, the form of study (full-time/ internal or external study mode), the form of payment (from the budget or from non-budgetary sources), student's citizenship, level of study (bachelor's, master's, specialty). The reason is that the grade is not an absolute indicator, it is just a relative indicator of academic performance. That is, when giving marks, teachers are guided by a certain average level of knowledge, which they consider as the level of assessment to be "good" and a certain level of knowledge that they consider unacceptable, that is, unsatisfactory, bad. In every institute of the university these levels are formed taking into account a number of special features. In this regard the division of students within the entire KFU into the so-called «successful» and «unsuccessful» seems to us unproductive.

In table 1 we indicate the average grades for institutions. They were obtained as the arithmetic mean from the marks in grades from 0 to 100 of all exams passed by those students who were taking part in the analysis program during the study period.

Table 1

Average grades (100-point grading scale) of the students of the KFU institutes

№	Institute	Average grade	Dispersion
	Technical institutes in total	81.7	0.48
1	Institute of Computational Mathematics and Information Technologies	82.7	0.46
2	Higher Institute of IT and Intelligent Systems	80.5	0.47
3	Institute of Mathematics and Mechanics	79.0	0.49
4	Institute of engineering	83.7	0.5
	Natural science institutes in total	81.7	0.44
5	Institute of chemistry	84.0	0.4
6	Institute of geology and petroleum technologies	80.5	0.42
7	Institute of Physics	80.1	0.45
8	Institute of Environmental Sciences	81.9	0.45
9	Institute of Fundamental Medicine and Biology	82.7	0.43
	Humanities institutes in total	85.8	0.39
10	Institute of Pedagogy and Psychology	88.5	0.29
11	Institute of International Relations	87.2	0.34
12	Institute of Philology and Intercultural Communication	87.4	0.35
13	Institute of Social and Philosophical Sciences and Mass Communications	86.0	0.36
14	Institute of Management, Economics and Finance	82.7	0.47
15	Faculty of Law	86.0	0.36

If we single out the best students in all KFU based only on the number of grades, a lot of students will be from the Institute of Psychology, the Institute of Philology, the Institute of International Relations. And a lot of students of the Institute of Mathematics and Mechanics, the Institute of Physics, the Institute of Geology will be among the worst. And the attempt to create a group of the best and worst students of KFU will present a group of the best students in the humanities and a group of the worst students of technical and natural science blocks.

The analysis of the distribution of average scores within institutes among students of different years and different fields of study showed that it also does not make sense to single out students in the category of good and bad academic performance by institution and by scientific fields. In some semesters students take fewer exams, in others – more. In the first years students adapt to the studying process at the university after school, there are a lot of students who are sent down from the university, etc.

As a result, we used the approach based on the selection of the best and worst students among those who are in the most similar conditions in every semester. It means that they study the same subjects and take the same exams, they work with the same teachers, in the same classrooms and buildings, they use the same equipment, etc. The selection of the best and worst students within student groups comply with these conditions.

If it is necessary to single out students with the evident high academic performance or, vice versa, with the evident bad academic performance, we single them out relative to the average indicator for the group. It is implied only in the average grade of students calculated in every semester.

The category of "average" students can be considered as the third category of students. In general, when looking for students with good academic performance, it is possible to use the approach when the entire group of students is divided only into academically successful and the rest. That is, for example, in a group of 20 students 10 people with the highest grades will be in the group of academically successful ones, and the other 10 students – to the group of the rest. Or 5 out of 20 will be classified as academically successful, and 15 – as the rest. But this approach seems rather unproductive. This is due to the fact that in many groups there is a high proportion of students with the same grades. And there will be the situation when one of the students with a certain average score will be defined as academically successful, and the another one – with the same grades – we will have to define as academically unsuccessful. In other words, it is necessary to single out a group of students with average academic performance, or the so-called average students.

Two approaches are possible in this connection. One approach is formal: in every academic group – regardless of the distribution of grades – we will assume that there are students with significant difference in grades and that we have the opportunity to divide them into 3 (academically successful, average and academically unsuccessful). In this situation it is important to understand what proportion of students will be left for further distribution into academically successful and unsuccessful students, that is, concerning the proportion of average students. The statement that the average grades of successful and unsuccessful students should differ greatly can be the only criterion for us.

And the number of successful and unsuccessful students can be taken as the same (for example, 25% of successful and 25% of unsuccessful students), as well as not the same (for example, 25% of successful and 50% of unsuccessful students with 25% average). These presumptions are based on the assumption that the assessment is not related to the real knowledge of students, that is, we compare the best and the worst students in the group, and we do not compare students with the objectively identified sample standard of knowledge.

On the basis of the above assumptions we can, for example, single out 40% of students as academically successful, 40% - as academically unsuccessful, and 20% - as the average ones in every group. Or we can single out 10% of students as the most successful students, 80% - as unsuccessful ones and 10% - as the average ones.

This approach to the selection of students in their groups into the most successful and unsuccessful ones is based on their selection from the entire mass of students on the formal basis of average grades in every semester.

The second approach to the division of students in their groups into academically successful, average and unsuccessful ones can be based on the search of actual differences between them in terms of grades. It means that if the range of grades in a student group is rather great, then it is possible to consider 40% of students as academically successful, 40% - as unsuccessful and 20% - as average students. But if in a student group the difference between grades is not great, then we will single out, for example, only 10% of academically successful students and 10% - as unsuccessful ones. And the major share will be consisted of average students. This approach is more time-consuming and it requires to analyze the distribution of grades in every academic group and to determine the number of students with great difference in grades and can be identified in opposite groups of academically successful and unsuccessful students.

We used exactly this variant of division of students into academically successful and unsuccessful ones. In every student group the students were ranked according to the average grade. Then in every group the same proportion of students (from 20% to 40%) was classified as academically successful and academically unsuccessful ones. This proportion depended on the distribution of the average student grade and on the number of students in the group. And respectively the proportion of students with average academic performance was from 60% to 20%. If in this or that group the average scores

of students were more or less the same, then only 20% of were considered as academically successful and 20% - as academically unsuccessful.

Supervised learning with a teacher would be one of the options to use neural networks in order to identify people with the potential for successful learning by aprofile in a social network. Metrics from the profile in the social network are the input data, and the learning with the materials of the training sample would be at the output. Such studies have already been carried out. Information about the results was presented in a number of publications [5], [6], [7].

This study made it possible to connect the interests of students and the level of their success with a probability of about 75% in the majority of institutes and university courses.

That is why we used a different approach. Unsupervised learning based on Kohonen neural networks was used.

We used numerical data from the profiles of students in the social network Vkontakte. Every object of this set is a vector xn=(x1, x2, x3, …, xn). In our study the dimension number of every vector from the remaining sample is 5. That is, each element of the vector represents a certain numerical profile indicator:

x1 – number of friends
x2 – number of subscribers
x3 – number of photos
x4 – number of videos
x5 – number of interesting pages.

This data was obtained using the API methods of the social network Vkontakte. In addition to the indicated numerical indicators, the following ones were also obtained from the VKontakte user profile: Number of posts on the wall, Number of gifts, Number of photo albums, Number of audio recordings. However, we did not use these parameters for two reasons. Firstly, a lot of users did not fill in these indicators. For example, only 20% of users had audio recordings. Secondly, the above-mentioned five indicators (that we selected for the analysis) had poor correlation with each other, that is, all indicators are very significant. Initially more than 50000 student profiles were obtained from the social network VKontakte. Despite the wide distribution of this network among students, it was impossible to use a significant part of the data obtained. Some students had several accounts of this social network, some students did not have an account at all, some students had a closed profile and data from their page could not be obtained, some students did not have any photos, videos and other parameters in their profile at all as far as they used their account in a social network as an analogue of Email. It is also necessary to say that there are a lot of foreign students at the university who use their traditional social networks, not Vkontakte. Also, there are the students who study on the Master's programs, who study on the external programs, there are students who are older than the majority of students. They use this social network very rarely and their profile does not contain any quantitative or qualitative information. We could judge about this by the indicator showing the last date of access to their page on the social network (this indicator was a year ago or even earlier) and by the indicator showing the degree of profile completion (it is close to zero). That is, we could not accept the profile data of such students as reflecting the qualities of students. As a result, out of about 50000 profiles found in the social network that could be associated with students, in the further research we were able to use quantitative data only from the previously mentioned 33484 students. Among them 8900 students were considered as academically successful and 8900 – as academically unsuccessful, that is, in total there were 17800 students. In the total amount their share was approximately 53%. The rest made up a group of students with average academic performance. The number and proportion of academically successful and unsuccessful students varied in different institutes.

Statistical processing of the data was carried out. It showed differences between groups of students with good and bad academic performance. This data is shown in Tables 2 and 3.

Table 2

Data that characterizes students with good academic performance

Indicator	Number of friends	Number of subscribers	Number of photos	Number of videos	Number of interesting pages
$\bar{X}$	239.9	290.4	82.4	137.1	98.7

D(X)	52790.1	138994.2	43593.7	183437.8	14136.4
Σ	229.7	372.8	208.7	428.2	118.8
A	3.7	12.7	9.5	10.1	6.2
E	23.4	274.5	151.8	159.4	98.2

where $\bar{X}$ – mean value, *D(X)* – dispersion, *Σ* – mean square deviation, *A* – asymmetry, *E* – kurtosis.

Table 3

Data that characterizes students with bad academic performance

Indicator	Number of friends	Number of subscribers	Number of photos	Number of videos	Number of interesting pages
$\bar{X}$	191.7	260.9	46.6	144.0	99.4
D(X)	66438.5	278433.1	20014.2	206657.9	18010.2
Σ	257.7	527.6	141.4	454.5	134.2
A	11.2	18.7	8.0	8.5	4.4
E	233.6	568.3	93.6	116.3	30.8

3. Results

We will present the results of using the Kohonen network for the analysis of the student profile data in the social network Vkontakte.

In order to carry out the clustering we used the Kohonen neural network which is able to break objects into clusters consisting of similar objects.

With its help a lot of students were divided into two, three and even four clusters.

In case of the neural network with two neurons, the following results were obtained:

– A lot of students were included into the first cluster: 64% - unsuccessful students, 36 % - successful students;

– The second cluster: 13% - unsuccessful students, 87 % - successful students.

That is, when dividing the set into two clusters we managed to recognize a group of successful students. The weights of the neurons – the centers of these clusters – were the following ones: 917.68; 755.94; 49.5; 1448.31; 283.46. These weights, in turn, are the coordinates of the neuron in space. The input vectors that were not farther than 88.95 are included in the second group with probability of 87%.

For the neural network with three neurons, the following results were obtained:

– The first cluster: 53% - unsuccessful students, 47 % - successful students.

– The second cluster: 12.5% - unsuccessful students, 87.5 % - successful students.

– The third cluster: 64% - unsuccessful students, 36 % - successful students.

That is, when dividing the set into three clusters we managed to recognize a group of academically successful students. The weights of the neurons – the centers of these clusters – are the following ones: 785.92; 907.89; 60.54; 1420.91; 506.68. The input vectors that were not farther than 8928 are included in the second cluster with probability of 87.5%.

For the neural network with four neurons, the following results were obtained:

– The first cluster: 11% - unsuccessful students, 89 % - successful students.

– The second cluster: 63% - unsuccessful students, 37 % - successful students.

– The third cluster: 56% - unsuccessful students, 44 % - successful students.

– The fourth cluster: 41% - unsuccessful students, 59 % - successful students.

That is, when dividing the set into four clusters we managed to recognize a group of academically successful students as well. The weights of the neurons – the centers of these clusters – are the following ones: 950.8; 767.2; 45.7; 1146.42; 453.6. The input vectors that were not farther than 89.18 are included in the first class with probability of 89%.

Based on the results of the Kohonen network we can say that in all cases the neural network managed to identify the group of academically successful students. However, Kohonen's network often fails to recognize academically unsuccessful students.

4. Conclusion

When deciding to hire a young specialist, it is very important to know how successful he was during the training period. Indirectly, this information can be obtained by analyzing a profile on a social network. In this regard, we collected the quantitative information in the social network Vkontakte (from the profiles of more than 50000 users), identified a set of quantitative parameters on the basis of which it is possible to do a study and to carry out experiments on clustering of data from students' profiles in the social network VKontakte. We selected the profiles of only those users who constantly log into this social network. Also, we did not use the data of social network users that did not meet all requirements.

We identified the relationship between the obtained clusters and students' performance. In the course of work all data was anonymous. A significant part of the work was devoted to the analysis of statistical data on the academic performance of the students at the Kazan Federal University and the determination of quantitative indicators of students' academic performance. We divided all students into three categories: academically successful students, academically unsuccessful students, students with average academic performance. Every student belongs to one of these student groups. Moreover, every group is determined among students who are in approximately equal learning conditions, studying the same subjects in the same period of time. Such conditions are created when teaching students in academic groups.

With the help of the Kohonen network we managed to cluster the users of the social network Vkontakte who are the university students. Statistical analysis does not allow to cluster the data of social network users. And for clustering we used poorly structured quantitative data on several metrics. These metrics were selected by us from a set of quantitative data based on the statistical analysis of the relationship between them.

5. Acknowledgements

The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation (Project No. 19-18-00253, "Neural network psychometric model of cognitive-behavioral predictors of life activity of a person on the basis of social networks").

6. References

[1] Feshchenko A., Goiko V., Mozhaeva G. et al. Analysis of user profiles in social networks to search for promising entrants // INTED2017 Proceedings, 11th International Technology, Education and Development Conference, March 6th-8th, 2017. – Valencia, Spain. – P. 5188–5194.

[2] Gafarov F, Vakhitov G, Enikeeva Z. Psychometric predictors of personal qualities for students of service and tourism department based on info from social networks // Journal of Advanced Research in Dynamical and Control Systems. - 2019. - Vol.11, Is.8 Special Issue. - P.2251-2255.

[3] Gafarov, F., Enikeeva, Z., Vakhitov, G., Nikolaev, K. The prediction of educational success of students-humanitarians in social networks from the psychometric view //International Journal of Pharmaceutical Research. - Volume 12, Issue 1, 2020, pp. 804-811.

[4] Galim Z. Vakhitov, Zulfira A. Enikeeva, Nadiya Yangirova, Adel Shavalieva, Pavel N. Ustin, Identification of the Clusters of Social Network Communities for Users with a Specific Characteristic // 12th International Conference on Developments in eSystems Engineering (DeSE) 7-10 Oct. 2019 (First Publication: 1 October 2019), pp. 775-784.

[5] Korshunov, A.V. (2013) Problems and methods for attribute detection of social network users. Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii [Electronic libraries: promising methods and technologies, electronic collections]. Proceedings of the conference. Yaroslavl: Yaroslavl State Technical University. pp. 380–390.
[6] Manzhula V.G., Fedyashov D.S., Kohonen neural networks and fuzzy neural networks in data mining // Technical sciences. - 2013. -N4. – pp. 108-114.
[7] Shastina A.E. Diagnostics of development of organizational and managerial competencies of the engineers using the self-organizing Kohonen maps // International Journal of Advanced Studies. - 2013. - T.3. - N4. - pp.28-33.