SOLVING WHEAT SELECTION PROBLEMS USING THE DBSCAN ALGORITHM

Shodiyev Fayzulla Yusupovich

Independent researcher, Digital technologies and artificial intelligence development research institute
ORCID: 0000-0001-7783-0502;
E-mail: fayzulloshyu@gmail.com

K E Y W O R D S

DBSCAN, legality, K-means, ε −neighboured, cluster, root object, boundary object, noise object, MinPts

A B S T R A C T

This article discusses the solution to the problem of identifying complex hidden relationships between varieties by clustering using the DBSCAN algorithm, dividing the features of soft wheat varieties into groups (reflecting the growth and development of the wheat plant, components of fertility and fertility, reflecting the quality of the grain).

Based on the values of the features related to different varieties of wheat, in one file created on the basis of experimental files, by dividing them into clusters and sets of noise objects using the DBSCAN algorithm, the problems of identifying wheat varieties that are close to each other in features and varieties that change their features due to abiotic factors, as well as sorting wheat varieties with low grain quality and low fertilities were solved.

The efficiency of clustering by dividing the features of objects into groups is shown using the example of wheat selection. ε −neighboured is determined, reflecting the degree of closeness of objects and the number of objects included in 3 groups of features related to wheat varieties. Clusters and noise objects are found based on the detected values. Based on this, it was possible to draw important conclusions about the wheat varieties that retained (did not retain) their bunch and were identified as noise objects in all 3 groups.

Introduction. By dividing the features of objects in the sample files into subgroups based on the level of similarity and performing clustering using the DBSCAN (Density-based spatial clustering of applications with noise) algorithm for each separated group, it is possible to effectively identify a number of complex hidden dependencies and legalities related to objects. The division of objects into subgroups is understood as groups divided according to the features of the feature (quantitative features (age, weight, height ...) and nominal features (color, profession, ...)), the units of measurement of the features (length (meter, centimeter, inch ...) [1], time (second, minute, hour ...)), the informativeness of the features, and the opinion of experts.

Applying the DBSCAN algorithm to each group leads to the following results:

Clusters - sets of objects that are similar in certain features;

Noise - objects that do not belong to any cluster and have their own features [2].

The article addresses the issue of identifying complex hidden legalities between varieties by clustering common wheat varieties according to their features.

Methods and materials. The idea of the DBSCAN algorithm used to solve the problem and the corresponding algorithm steps are presented.

The advantage of this algorithm over the K-means [3,4] clustering algorithm is that the DBSCAN algorithm is able to select clusters of arbitrary shape, and in addition, it determines the number of clusters in the sample file itself. Although this algorithm was first proposed in 1996, it is still widely used today. It is a purely algorithmic [5] approach that is not directly related to probability theory and the density of the data distribution.

The idea of the algorithm starts from the concept of the ε −neighboured of an object. For an arbitrary vector x in a metric space, the set of points lying in a region not larger than the neighborhood ε is defined as follows:

©F.Shodiyev

$$U_\varepsilon(x) = \{u \in U : \rho(x,u) \le \varepsilon\} \qquad (1)$$

where $\rho(x,u)$ is the chosen metric for the space of symbols. For example, the Euclidean metric [6]. The value of $\varepsilon(\varepsilon > 0)$ is input during the operation of the DBSCAN algorithm.

Then, based on the ε −neighboured value, objects are divided into three categories:

1. Root objects: m objects $|U_\varepsilon(x)| \ge m$ belonging to the ε −neighboured.
2. Boundary objects: objects that are not roots, but lie on the boundary of the ε −neighboured;
3. Noise objects: objects that have neither a root nor a boundary.

From this it can be seen that this algorithm is based on heuristics, not mathematics [7]. Object types are a form of heuristics.

Suppose we are given a data set consisting of a two-dimensional feature space. First, we randomly select an object x_i from this set. If this object x_i has fewer than m root objects in its ε −neighboured, then this object is considered noise. Then, the next object x_i is randomly selected from the remaining objects [8].

If m root objects are found lying in the ε −neighboured of the selected object x_i, then the set of found objects is defined as the root vector and the above processes are repeated recursively. In addition, if the object does not have a sufficient number of neighbors in the ε −neighboured, it is defined as a boundary object, otherwise as a root object. As a result, we consider all objects belonging to the ε −neighboured. Clustering is performed in this way [9].

Again, this process starts anew, objects that are included in the cluster or identified as noise do not participate in this process. In this process, an object x_i is also randomly selected and it forms a noise object or cluster. As a result, after examining all the objects in the sample [10], we have a set of objects divided into clusters and identified as noise objects. In addition, the number of clusters is automatically determined based on the given parameters ε and m.

DBSCAN algorithm execution steps:

1. Reading objects and their corresponding attribute values from a sample.
2. Enter the value of ε (epsilon). This value is a radius (distance) and is used to identify objects close to the selected object. That is, if the distance between two objects is less than ε, they are considered "near" objects to each other.
3. MinPts. This parameter specifies the minimum number of objects that must be selected and nearby. If the number of objects nearby the selected object is greater than or equal to MinPts, it is called a "core" object.
4. Identification of "core" objects. $CorePoint : \{\rho \in D || N_\varepsilon(\rho)| \ge MinPts\}$, where $N_\varepsilon(\rho)$ is the set of neighboring objects within radius ε of object ρ.
5. Identifying neighboring objects. If object ρ is a core object, then all objects around it (i.e., $N_\varepsilon(\rho)$) are also neighbors to this object, and they are grouped together in a single cluster: $ClusteredPoints = \{\rho | N_\varepsilon(\rho) \ge MinPts\}$.
6. Clustering. If an object is a core object, all neighboring objects around it (core or boundary) are combined into a single cluster. In this way, the belonging of each object to a cluster is determined. This process continues recursively, and each neighboring object checks other objects in its surroundings, and if they are neighbors, they are added to the cluster.
7. Identifying peripheral (noise) objects. An object is considered a peripheral object if it is not a core object or if it is not close to any core object:$NoisePoints = \{\rho || N_\varepsilon(\rho)| < MinPts\}$. Such objects are not included in the formed clusters.

Results and discussions. Below, the features of the objects in the training sample are divided into small groups based on the level of similarity and each divided group is divided into clusters using the DBSCAN algorithm. In this way, it is possible to effectively identify a number of complex hidden relationships and legalities related to the objects.

A number of features related to wheat varieties are divided into the following groups:

(cc) BY ©F.Shodiyev

1. A group of features reflecting the growth and development (morphological and phenological features) of a wheat plant:

- vegetation period;
- plant height;
- terminal culm length;
- spike length.

2. Group of features related to productivity and crop components:

- number of ears;
- fertility;
- 1000 grain weight;
- grain type.

3. Group of features reflecting grain quality (nutritional and technological properties):

- protein content;
- gluten content;
- gluten deformation index (GDI).

Applying the DBSCAN algorithm to each group leads to the following results:

Experiment 1. According to the group of features reflecting the growth and development of wheat plants, $\varepsilon = 4.9$, with *MinPts=3*, the wheat varieties in the training sample were divided into 9 clusters and 65 noise objects (Figure 1).

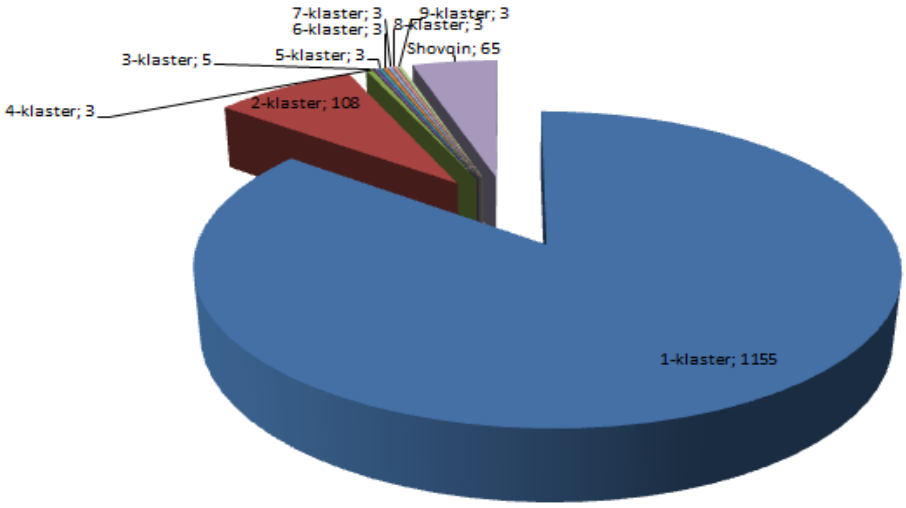


Figure 1. *Distribution of wheat varieties into clusters according to morphological and phenological features.*

Experiment 2. According to the group of traits related to fertility and fertility components of wheat varieties, $\varepsilon = 8.3$, with *MinPts=3*, the wheat varieties in the training sample were divided into 3 clusters and 66 noise objects (Figure 2).

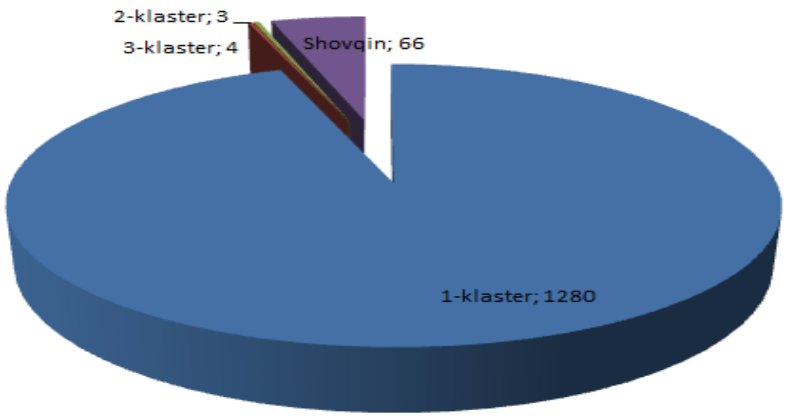


Figure 2. Distribution of wheat varieties into clusters according to fertility and fertility component traits.

©F.Shodiyev

Experiment 3. According to the group of traits reflecting grain quality, $\varepsilon = 2.3$, with *MinPts=3*, the wheat varieties in the training sample were divided into 8 clusters and 63 noise objects (Figure 3).

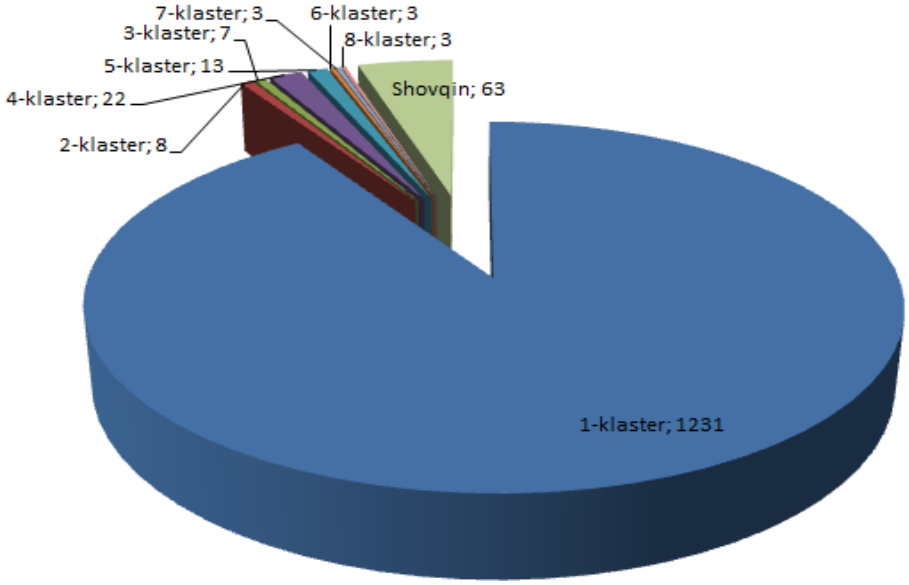


Figure 3. *Distribution of wheat varieties into clusters by group of traits reflecting grain quality.*

From the results obtained, it can be determined that there are correlations between the groups presented. For example, according to industry experts, the duration of the "Vegetation period" affects the "Fertility" feature. That is, in years with a short "Vegetation period", high yields of early-ripening varieties or low yields of late-ripening varieties were observed. In addition, it was found that the "Fertility" feature, in turn, affects the features of protein and gluten content, and a complex system of such correlations was also found using the DBSCAN algorithm.

Discussions. Applying the DBSCAN algorithm to the above selection work significantly increases the efficiency of the work. For example, wheat varieties that have maintained or changed their cluster in different years can be easily identified according to the results obtained from the algorithm. Varieties such as Bunyodkor, G'ozg'on, Grom have not changed their cluster for several years. This is shown by the results obtained based on the algorithm. Some varieties have been identified as "Noise" objects, since their characteristics have changed over the years.

In particular, the results obtained using the algorithm have proven that it is more effective to perform the clustering process according to groups of features. Also, by comparing the objects located in the clusters and considered as noise in each group to reach the final conclusions, it became easier to make final decisions on varieties whose place is in doubt.

In this way, it is also possible to easily distinguish varieties and ridges that have maintained (or changed) their characteristics for several years under different weather conditions.

Conclusion. The effectiveness of clustering the features of objects (varieties and ridges) into groups in wheat selection was demonstrated. The optimal numerical values representing the degree of proximity of objects were determined for the features of wheat varieties, reflecting the morphological and phenological features of the wheat plant, the features related to fertility and crop components, and the features reflecting the nutritional and technological properties of wheat. Based on the determined values, clusters were formed and noise objects were identified.

In all three groups, it was possible to draw important conclusions about the wheat varieties that retained (or did not retain) their cluster and were identified as noise objects.

(cc) BY ©F.Shodiyev

References:

1. Jörg Sander, Martin Ester, Hans-Peter Kriegel, Xiaowei Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications // Data Mining and Knowledge Discovery. — Berlin: Springer-Verlag, 1998. — Т. 2, вып. 2. — С. 169–194.
2. Игнатьев Н. А., Згуральская Е. Н., Марковцева М. В. Нелинейные преобразования признаков и поиск закономерностей на данных больных хроническим лимфолейкозом //Информационные технологии и нанотехнологии (ИТНТ-2020). Сборник трудов по материалам VI Международной конференции и молодежной школы (г. Самара, 26-29 мая). – 2020. – №. 4. – С. 123-128.
3. Шодиев Ф., Эшбоев Е., Суярова А. Прогнозирование устойчивости к болезням высококачественных сортов пшеницы с использованием метода расчета обобщенных оценок //E3S Web of Conferences. – EDP Sciences, 2023. – Т. 401. – С. 04063.
4. Klicheva F.G., Eshboyev E.A. (2023) Creation of an intelligent system to support medical diagnosis. Innovative technologica-methodical research journal. Volume 4, Issue 5 May 2023, 82-87.
5. Shodiyev F. Intellectual system based on the determination of hidden legality //Central Asian journal of education and computer sciences (CAJECS). – 2022. – Т. 1. – №. 5. – С. 11-16.
6. Мадрахимов Ш. Ф. Отбор шумовых объектов на базе обобщённых оценок //Проблемы вычислительной и прикладной математики. – 2018. – №. 2. – С. 122-131.
7. Fayzulla S., Munisa D. DETERMINATION OF INFORMATIVE FEATURES USING THE METHOD OF DIVISION INTO INTERVALS BASED ON THE COMPACTITY HYPOTHESIS //Universum: технические науки. – 2024. – Т. 7. – №. 3 (120). – С. 25-29.
8. Мадрахимов Ш. Ф., Саидов Д. Ю. Группировка признаков по критерию устойчивости объектов классов //Актуальные проблемы прикладной математики, информатики и механики. – 2016. – С. 93-95.
9. Игнатьев Н. А., Згуральская Е. Н., Марковцева М. В. Поиск скрытых закономерностей, влияющих на общую выживаемость больных, методами интеллектуального анализа данных //Искусственный интеллект и принятие решений. – 2020. – №. 3. – С. 73-80.
10. Дилмуродов, С., Мейлиев, А., Джураев, Д., Бойсунов, Н., Хазраткулова, С., Шодиев, Ф., ... & Абдимаджидов, Д. (2025). Оценка генетической изменчивости и биометрических показателей у сортов мягкой пшеницы: значение для современных методов селекции. В BIO Web of Conferences (Vol. 163, p. 03001). EDP Наука.

©F.Shodiyev