UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH W BYDGOSZCZY
ZESZYTY NAUKOWE NR 270
TELEKOMUNIKACJA I ELEKTRONIKA 23 (2019) 23–32

GWAS DATA ANALYSIS WITH THE USE OF MACHINE LEARNING ALGORITHMS – REVIEW

Sylwester Michał Kloska[1], Anna Marciniak[1,2]

[1] Nicolaus Copernicus University Ludwik Rydygier Collegium Medicum in Bydgoszcz, Department of Medicine, ul. Jagiellońska 13-15, 85-067 Bydgoszcz, Poland

[2] University of Science and Technology, Faculty of Telecommunications, Computer Science and Electrical Engineering, Al. Prof. S. Kaliskiego 7, 85-796 Bydgoszcz, Poland

Summary: Machine learning is a part of field concerned with AI. The main goal of machine learning algorithms is to create automatic system that improves itself with the use of its experience (given data) to gain new knowledge. Genome-Wide Association Studies compare whole genomes of different individuals in order to see if any of genetic variants are correlated with a trait. Using ML for GWAS analysis can be beneficial for scientists. It has been proved several times in various ways.

Keywords: machine learning, genome-wide association studies, GWAS, artificial intelligence, bioinformatics

1. INTRODUCTION

Machine learning (ML) is a part of field concerned with AI (artificial intelligence). It is interdisciplinary study that involves informatics, robotics and statistics. The main goal of machine learning algorithms is to create automatic system that improves itself with the use of its experience (given data) to gain new knowledge. Machine learning is in fact a baby of AI studies and tries of its practical use. ML is most commonly used in the development of new technologies and industry. Special algorithms should allow software to automate data gathering and data analysis in order to perfect itself. Gained experience should allow the system to process data and achieve similar results more effective and faster. There are several advantages of ML use in many fields of everyday life, such as increasing effectiveness and reliability, as well as lowering the cost.

One of the first models of ML is the project of Arthur Samuel, an employee of IBM, that in 1952–1962 was improving the system that was used in training of chess players. One of the biggest breakthroughs of AI and ML studies was creation of expert system Dendral on Stanford's University in 1965. It was created to automate analysis and identification of chemical molecules unknown to chemists back then. Results obtained with the use of Dendral system were first that were published in a scientific journal. The interest in ML had steadily increased. In early 1990's, Gerald Tesauro created a program that was able to compete with world champions in a board game Backgammon. To achieve this level of perfection the program was analyzing over a million of its own games. Later on, Tesauro's brilliant program was adopted in neuroscience. The main goal was to use ML in practical problem solving. It was also beneficial for ML that data became more digitalized in 1990's and it was possible to distribute the data via Internet [15].

Creators of Deep Blue computer were so confident of its abilities that they even challenged Garri Kasparow, famous chess player and a world champion in chess, widely recognized as the greatest chess player of all time. Their program was able to win with Kasparow. It proves the potential ML has and what can be achieved with the use of it.

ML in theory is able to create new concepts, detect unknown regularities in given data, formulate decision rules, assimilate new concepts and structures through generalization and analogy, modify, generalize and specify data, acquire knowledge through interaction with the environment, formulate human-understandable knowledge. For this reason ML can be used in various fields, for example:

- analysis and use of huge databases. Size of these databases as well as their complexity and requirement of continuous updating prevent non-automated analysis (e.g. in such areas as economics, medicine, chemistry);
- adaptation of the system to the environment through dynamic modification. That allows proper operation in changing conditions (e.g. robotics, control systems, production, data analysis).;
- in searching and analyzing dependencies in large databases in order to present information synthetically according to given criteria (e.g. expert systems, internet search engines).

ML is constantly evolving and finding its use in various fields. The number of possible use of ML is enormous and it is predicted that in the future every aspect of technique will include ML algorithms. For now it is used for voice recognition, automatic navigation and data analysis and classification.

Despite fast development and the growth of interest in ML there are also risks, problems and limitations. One of them is the fact that they are still human-dependent. Creation process relies on man – human is responsible for describing the algorithm how it should gather the data, how to analyze and use them. But the creation is not the only problem concerned with ML. Other issues are:

- too low or too high system dependence on the environment in which it is located, which may lead to incomplete data analysis or misinterpretation,
- credibility and correctness of the conclusions generated and inductive reasoning cannot be fully proved, but only falsified,
- incomplete or partly contradictory data,
- not defining domain restrictions may lead to far-reaching generalizations and erroneous conclusions.

These problems led to adoption of the following postulates:

- knowledge generated by the system should be subject to human control and assessment, according to the criteria provided by him,
- the system should be able to provide an explanation in the event of a problem,
- knowledge should be understandable to man, i.e. expressive in the description and mental model adopted by him.

As previously mentioned, ML heavily depends on dataset. It uses training data in order to find certain properties of the data and then use it to predict outputs. Algorithm analyzes the data, both input and output. Based on that, a mathematical model is created and it is further used to find patterns in the data and to predict the outcomes based on inputs. Input data can vary, it can be numeric data, figures, even sound patterns. More specialized algorithms can be effective even when the input data is incomplete – only partially available. One widely known algorithm is used in classification of incoming emails. It filters email traits in order to allocate the email in a proper folder [2].

Creating a model is necessary to perform machine learning. The model is trained on training data and then can be used to process additional data and make predictions. There are several types of models used in ML: artificial neural networks (ANN), decision trees, support vector machines (SVM), Bayesian networks (BNN), genetic algorithms (GA) [8].

The amount of training dataset is very important matter in order to create a good algorithm. On the one hand, the bigger training data is, the higher chances are for algorithm to predict outputs correctly, but on the other hand it is important to not overtrain the algorithm, also known as overfitting.

Despite ML great success in a variety of fields, such as bioinformatics, banking, marketing, medical diagnosis and many more, it is also possible that ML fails and not deliver expected results. There is a great variety of reasons for that, like too little data, data bias, poorly chosen tasks and algorithms and evaluation problems. However, previously mentioned problems are not holding back people responsible for development of new and better models. There is still great hope for machine learning as it is gaining popularity in genetics [16].

Genome-Wide Association Studies (GWAS) is a study that compares whole genomes of different individuals in order to see if any of genetic variants are correlated with a trait (Fig. 1). GWAS mainly focuses on associations of single-nucleotide polymorphisms (SNPs) and traits like human diseases, but it can be also applied to any other genetic variants and any other organisms. Studies on human datasets compare DNA of participants with their phenotype traits or diseases. If one of the variants (allele) is present more frequently in people with certain trait or disease, then the variant is called as "associated" with the disease. It does not always mean it is the marker of a disease, but can be due to various genetic and non-genetic reasons.

Since GWAS examine whole genome, it is called *non-candidate-driven* approach. It is opposite to the *gene-specific candidate-driven studies*, which focus on certain genome region in order to find a relationship with an existing feature in this specific region. GWAS can be used to find SNPs and other DNA variants that are associated with a disease, but they cannot be used to confirm which of the genes are causing the disease. Until now more than 3000 GWAS have been performed and studied over 1800 traits and diseases. It contributed to finding of thousands of associated SNPs. Most of them may be pretty weak, but there are some of major importance regarding rare genetic disease [13, 14].

To begin with, its noteworthy to point out that DNA samples of two different individuals differ in millions of ways. Genomic difference can be in single nucleotides (SNPs) as well as larger DNA regions variations, like insertions, deletions (In/Del) and copy number variations (Fig. 2). All of them affect the fact that these individuals exhibit different phenotypic features – they look different, they may have different diseases, or are in risk of different diseases. Before GWAS, it was most common to study family members and check their medical history to find out genetic associations. However, this approach did not give reproducible results in case of complex diseases. For this reason, genetics thought that GWAS could be useful as a diagnostic tool. Early statistical calculations proved that GWAS may be better than linkage studies. It was also beneficial for GWAS that genetic tools were on the rise back then and more genetic study methods became more available and results they gave were reliable.

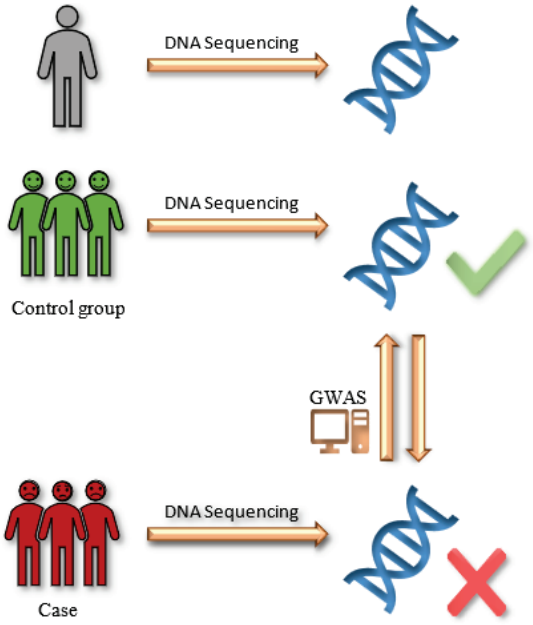

Fig. 1. Schematic illustration of GWAS. After DNA sequencing results obtained from control group and case are compared to find associations with specific case traits

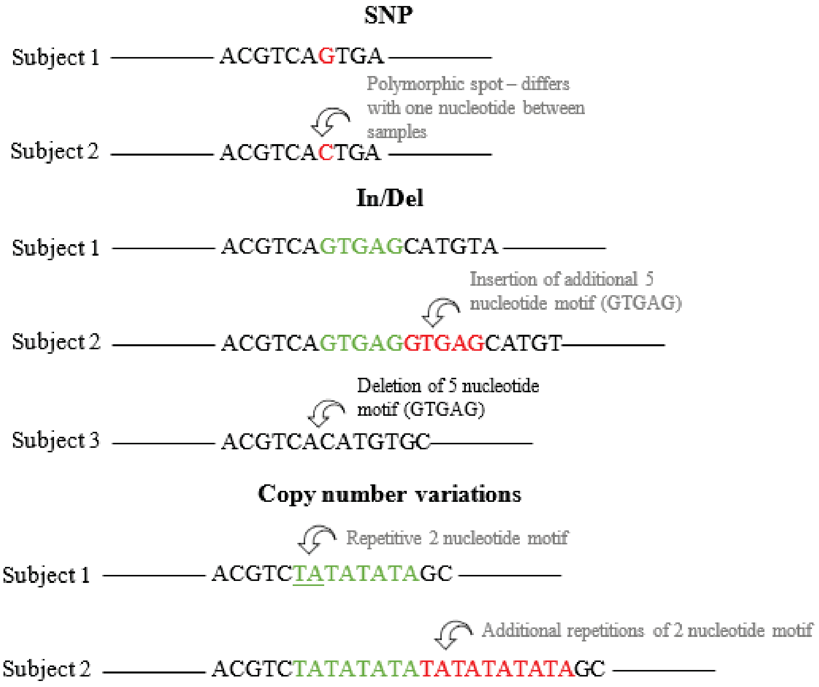
SNP
Subject 1
ACGTCAGTGA
Polymorphic spot – differs with one nucleotide between samples
Subject 2
ACGTCACTGA
In/Del
Subject 1
ACGTCAGTGAGCATGTA
Insertion of additional 5 nucleotide motif (GTGAG)
Subject 2
ACGTCAGTGAGGTGAGCATGT
Deletion of 5 nucleotide motif (GTGAG)
Subject 3
ACGTCACATGTGC
Copy number variations
Repetitive 2 nucleotide motif
Subject 1
ACGTCTATATATATAGC
Additional repetitions of 2 nucleotide motif
Subject 2
ACGTCTATATATATATATATATATAGC

Fig. 2. Graphic representation of most common basic genetic variations: SNP, In/Del, Copy number variations

Most commonly used strategy in performing GWAS bases on comparing two groups of individuals – people from one group are healthy (they do not indicate certain trait – control) and in the other one there are people with a disease (case). Genotyping is performed for all of participants and usually more than millions of SNPs are found. Then the SNP are examined between control and case groups, to check if their frequency in these two groups differs significantly to associate them with a trait. Statistical methods are used at this stage of study [14].

At the beginning genome-wide studies focused mostly on contribution of single SNP, however further genetic test and expansion of knowledge suggested this approach was not fully correct. As it is known now, traits may depend on more than one SNP. Nowadays scientist try to combine GWAS with the data achieved in protein studies in order to obtain more detailed information. The future challenge for GWAS is to use obtained results in order to accelerate drug and diagnostics methods development [3].

This is where both GWAS and machine learning meet (Fig. 3). Data obtained by genotyping with the use of NGS (Next Generation Sequencing) is flooding geneticists. Human genome consists of more than 3 billion base pairs. However, NGS methods are delivering even more data, so they can be transitioned into genome sequence later on, with the use of special algorithms. The amount of data is too large for scientists to examine them on their own, this is where machine learning could be used as an useful tool. Several algorithms have been optimized to find SNPs located throughout genomes and correlate them with traits, based on given data. This can be helpful in discovering new disease-associated SNPs.

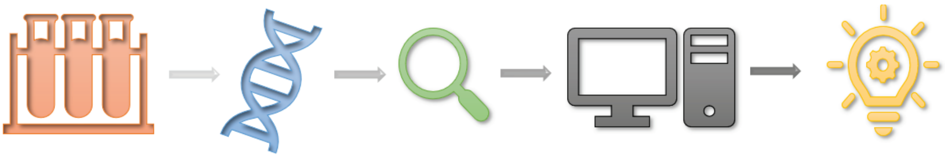

Fig. 3. Graphic representation of the presented problem – from DNA isolation, through analysis of its nucleotide sequence, to the use of bioinformatic knowledge and machine learning methods

Large datasets are causing problems. There are a lot of genome-wide studies in progress. All of them are delivering databases with obtained results, however, these databases are not suitable for combining with others. They may have different structure, e.g. different column assignment etc. This is why many scientists are still obligated to transcript the results manually. This task becomes nearly impossible, knowing how many of databases are appearing and providing new datasets. This is why a new tool for analyzing the genetic variants is needed. It would be helpful to create a specialized algorithm, that allows extracting of the most important information and to prescribe it in a universal way, so it can be accessed later on with ease. This is quite demanding thing to do, and requires a lot of preparation.

2. MACHINE LEARNING IN BIOINFORMATICS

2.1. Diagnosis of diseases

Diseases that are responsible for the most of deaths worldwide, such as various types of cancer, cardiovascular disease, neurodegenerative diseases are caused by both environmental and genetic factors. Most of them are not the results of a single mutation, but rather genetic changes in many various genes. This is why it is so important to obtain as much information as possible from genome-wide studies, to understand the mechanisms of disease. There are beliefs that ethnicity may affect human genetic profiles. Nikoghosyan and her team used model based on self-organizing maps (SOM) ML. It has been previously used with success in other genetics and bioinformatics areas. For this reason it was chosen as a tool to study SNPs associated with diseases. They examined around 44,000 SNPs across 52 populations. Results of these research gave so called "SNP portraits" that were dependent on populations. Obtained results demonstrated that some populations can be extremely genetically diverse. Observations helped to find out that different populations may have different predispositions to certain diseases. Such results support the presumption of multifactorial disease base [12].

Khan *et al.* raised the issue of genes that could potentially affect the development of mental disorders. For this purpose, the iMEGES (integrated mental-disorder genome score) tool was developed to analyze the entire genome/exome sequence, and then, using a deep neural network on TensorFlow framework to check the variants obtained for those affecting mental illness. In this case, input data is genetic mutations and phenotypic information from a patient that suffers from mental disorder, and output data is rank of whole genome sensibility variants and the prioritized disease-specific genes for mental disorders. This tool has been tested on various datasets, e.g. subjects with schizophrenia and autism spectrum disorder. By using iMEGES it is possible to reduce the susceptibility of people to mental illness as well as constructing personalized therapy [7].

There have been research that tested the use of ML in order to find out correlation between genetic variants and lipid traits, low-density lipoproteins cholesterol (LDL-C), high-density lipoproteins cholesterol (HDL-C) and triglycerides (TG) levels. There were a lot of studies that provided reliable information confirming these lipid traits to have high impact on cardiovascular disease risk. However, genetic background for lipid levels is not well understood. Mentioned model was used to prioritize genetic variants to correlate it with lipid traits [9].

2.2. Creating SNP based models

Merelli *et al.* noticed high potential in analyzing SNPs, as it is a great source of information how genetic variants may affect phenotype traits. Genetic methods used nowadays allow quick and effective analysis of 1 million SNPs, targeting those which are known to be associated with diseases and traits. They also noted that it is worth using information obtained by other researchers before. This is how they got an idea of SNPranker 2.0. This tool was created to prioritize SNPs, by features that are in a particular interest of its user, such as epigenetics and functional genomics attributes. SNPranker 2.0 is an algorithm, that relies on machine learning. It was optimized with the use of experimental results. A genetic algorithm was created to find SNPs related to an input dataset of genes and biological processes. As a result, SNPranker 2.0 provides a list of SNPs together with statistical probability of the most presented pathologies. They chose supervised ML to create a function, that was able to connect inputs and outputs. This choice demanded a carefully prepared set of training and validation datasets, because the model is created basing on this data. It was crucial to find optimal weights for various features, as it ensures the best possible sensitivity and specificity. SNPranker 2.0 was tested and approved as an useful tool for SNP prioritization [11].

It is possible that certain SNPs do not affect the trait alone, but rather in a combination with other SNPs. It is called epistasis. Detection of these epistatic SNPs can be very difficult and tricky task. However, a correct detection can be used to improve prevention, diagnosis and diseases treatment. Designing a powerful method to identify epistatic interactions between SNPs is a big challenge for bioinformatics. It is demanding and difficult considering the size of data and the large amount of combinations between genetic factors. Han *et al.* created DASSO-MB algorithm, that has its background on Markov Blanket method. The aim of this algorithm is to find epistatic interactions in GWAS results. They ensure that the algorithm they created detects SNPs have high association with diseases, but also it generates very few false-positive results. Their algorithm, DASSO-MB, uses a heuristic search strategy. It is calculating the association between variables in order to avoid the time-consuming training process, which takes place in other machine-learning methods. It was tested both on simulated and real datasets and it met expectations. Their study also indicated superiority of DASSO-MB in comparison with other algorithms. They noted that GWAS generates a lot of data, and it is necessary to save potential costs of biological experiments and to be as effective as possible in pathogenesis research [5].

2.3. Determining ethnicity and ancestors

One of research teams propose a novel use of machine learning method called "ETHNOPRED". This algorithm uses disjoint decision trees to predict an individual's both, continental and sub-continental ancestry. This project uses genotype and ethnicity data from HapMap project. In order to predict individual's continental ancestry ETHNOPRED created an ensemble of 3 decision trees consisting of total 10 SNPs, with 10-fold cross validation accuracy of 100% using HapMap II dataset. After extending this model to 29 disjoint decision trees over 149 SNPs (some of SNPs values were missing in samples) it was possible to achieved accuracy of ≥ 99.9%. ETHNOPRED was also tested on independent dataset of Caucasian origin, where accuracy was 96.8%. In order to learn classifiers to distinguish subpopulations HapMap III dataset was used. To do so ensembles of 3, 11, 21, 25 and 39 disjoint decision trees involving various number of SNPs were used. In this case accuracy was: in worst case 86.5% ± 2.4% and in best case 98.3% ± 2.0% (Table 1). To sum up, ETHNOPRED is a novel technique that uses decision trees for producing classifiers that are able to identify an individual's continental and sub-continental heritage. Decision trees were used due to the ease of use, short training time and results that are easy to interpret. ETHNOPRED uses small amount of SNPs and produces high accuracy results [4].

Table 1. ETHNOPRED accuracy including subpopulations and the number of SNPs analyzed

Subpopulation	Disjoint decision tree ensemble	Number of SNPs	Accuracy
European	3	31	86.5% ± 2.4%
East Asian	39	502	95.6% ± 3.9%
African	21	526	95.6% ± 2.1%
North American	11	242	98.3% ± 2.0%
Kenyan	25	271	95.9% ± 1.5%

Jain *et al.* used the data regarding ethnicity from more than 950 articles as training data, and then tested the algorithm with results from 307 articles. They accomplished accuracy at around 90% accuracy for identifying target trait mention of a GWAS study. This results is quite satisfying which proves that it is effective approach, however, this result still leaves room for improvement [6].

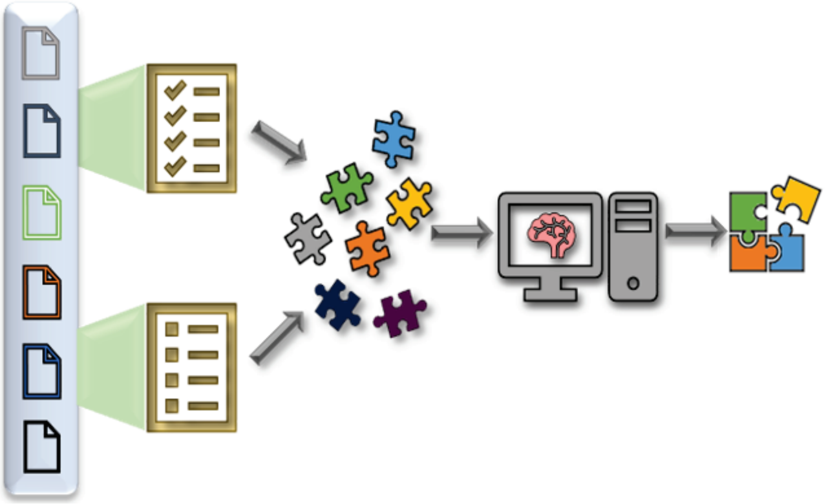

Fig. 4. Graphic summarization of machine learning. Standard approach consisting in selection of the information for ML training process in order to obtain the best possible outcome

2.4. Bacterial diseases

The issue of machine learning in the context of genomic data also became an interest of Long and his associates. In order to find genetic determinants of various phenotypic traits they implemented two algorithms: 1) adaptive boosting (AB) – which is considered as one of the best classifiers, 2) repeated random forest (RRF) – a modified version of RF. These algorithms had to facilitate data analysis. As another example of supervised ML, they were trained on labelled train datasets. To assess functionality of these models Long and his team used data from influenza, as it is relatively easy to determine three phenotype traits. Those traits were infectivity, transmissibility, and pathogenicity, and they were known from experimental evidence. The team performed sensitivity tests and obtained even 100% correct predictions basing on just 20 sequences. However, we should keep that in mind, that influenza has a small genome. Their predictions suggest that sensitivity would drop to around 90% in case of organisms with bigger genome and also their model would need more sequence data for these bigger genomes. Next step of their research was to predict the drug resistance determinants to Ciprofloxacin, Ceftazidime and Gentamicin, in a bacterium *Pseudomonas aeruginosa*. Results obtained on influenza datasets with both algorithms were satisfying. However, their prediction abilities were worse in case of bacterial data, but even on this data RRF performed slightly better than AB. It is also noteworthy, that these algorithms were unbiased, they did not have any assumptions before. They were just using given datasets. The team has proved that ML algorithms can be used in genetic determinants prediction, even when we are in disposal of data limited only to few items [10].

Beam *et al.* created BNN (Bayesian Neural Network) and its performance was tested on data obtained from GWAS designed to look for genetic markers associated with tuberculosis (TB). Dataset carried 60,000 SNPs from 105 subjects. Each person was classified into 1 of 2 groups: 1) people who were currently infected with any active form of tuberculosis, 2) people having latent from TB. It was considered that BNN works better than MDR compared with it, because MDR is not suitable for such large amounts of data. Performed studies demonstrated that BNN can be used as a powerful tool for analyzing association studies, as it has capability of large datasets obtained from GWAS [1].

3. CONCLUSIONS

The above review indicates the great potential of using machine learning in genetic research and analysis. The use of algorithms to obtain reliable results is extremely helpful for scientists, due to the amount of data being buried in connection with the ever-increasing popularity of GWAS research. As presented above, GWAS can be used for many purposes – searching for the genetic basis of diseases, genetic testing of ethnic origin and many others. The use of ML significantly speeds up this type of analysis, and its effectiveness reaches a satisfactory level. The increase in interest in bioinformatics in recent years indicates the need to automate the evaluation of results – machine learning methods may be able to "see" and adapt to regularities invisible to people involved in the analysis of results.

BIBLIOGRAPHY

[1] Beam A.L., Motsinger-Reif A., Doyle J., 2014. Bayesian neural networks for detecting epistasis in genetic association studies. BMC Bioinformatics 15(1), 1–12.

[2] Bishop C., 2006. Pattern Recognition and Machine Learning. Springer Cambridge.

[3] Bush W.S., Moore J.H., 2012. Genome-Wide Association Studies (Chapter 11). PLoS Comput. Biol 8(12).

[4] Hajiloo M., *et al.*, 2013. ETHNOPRED: A novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. BMC Bioinformatics 14.

[5] Han B., Park M., Chen X.-W., 2010. A Markov blanket-based method for detecting causal SNPs in GWAS. BMC Bioinformatics 11(SUPPL. 3).

[6] Jain S. *et al.*, 2016. Weakly supervised learning of biomedical information extraction from curated data. BMC Bioinformatics 17(1), 1–12.

[7] Khan A., Liu Q., Wang K., 2018. iMEGES: Integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes. BMC Bioinformatics 19(Suppl. 17).

[8] Larranaga P. et al., 2005. Machine learning in bioinformatics. Brief. Bioinform 7(1), 112.

[9] Leal L.G. *et al.*, 2019. Identification of disease-associated loci using machine learning for genotype and network data integration. Bioinformatics May, 1–9.

[10] Long G.S., Hussen M., Dench J., Aris-Brosou S., 2019. Identifying genetic determinants of complex phenotypes from whole genome sequence data. BMC Genomics 20(1), 1–17.

[11] Merelli I., *et al.*, 2013. SNPranker 2.0: A gene-centric data mining tool for diseases associated SNP prioritization in GWAS. BMC Bioinformatics 14(SUPPL.1), 1–12.

[12] Nikoghosyan M., Hakobyan S., Hovhannisyan A., Loeffler-Wirth H., Binder H., Arakelyan A., 2019. Population levels assessment of the distribution of disease-associated variants with emphasis on Armenians – A machine learning approach. Front. Genet 10(APR), 1–16.

[13] Pandey J.P., 2010. Genomewide association studies and assessment of risk of disease. N. Engl. J. Med. 363(21), 2076–2077.

[14] Pearson T.A., Manolio T.A., 2008. How to Interpret a Genome-wide Association Study. JAMA 299(11), 1335–1344.

[15] Samuel A.L., 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. Der. 3(3), 210–229.

[16] Tarca A.L., Carey V.J., Chen X.-W., Romero R., Drăghici S., 2007. Machine learning and its applications to biology. PLoS Comput. Biol. 3(6).

ANALIZA DANYCH GWAS PRZY UŻYCIU ALGORYTMÓW UCZENIA MASZYNOWEGO – PRZEGLĄD LITERATURY

Streszczenie

Uczenie maszynowe jest dziedziną nauki związaną ze sztuczną inteligencją. Głównym celem algorytmów uczenia maszynowego jest stworzenie automatycznego systemu, który poprawia się dzięki wykorzystaniu swojego doświadczenia (danych) w celu zdobycia nowej wiedzy. Badania asocjacyjne całego genomu (GWAS) porównują całe genomy różnych osobników, aby sprawdzić, czy którykolwiek z wariantów genetycznych jest skorelowany z cechą. Wykorzystanie ML do analizy GWAS może być korzystne dla naukowców. Zostało to udowodnione na różne sposoby.

Słowa kluczowe: uczenie maszynowe, badania asocjacyjne całego genomu, GWAS, sztuczna inteligencja, bioinformatyka