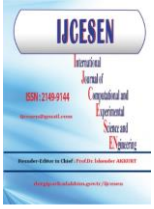
International Journal of Computational and Experimental Science and Engineering (IJCESEN)
Vol. 4-No.2 (2018) pp. 23-31
http://dergipark.gov.tr/ijcesen



ISSN: 2149-9144

Research Article

Big data Analysis in Plant Science and Machine Learning Tool Applications in Genomics and Proteomics

Rana VALİZADEH-KAMRAN[1], Ahmad HEYDARİYAN[2], Najmeh DAMGHANİ[3], Jalil NOURMOHAMMADİ-KHİARAK*[4]

[1]Department of Biotechnology, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran.
[2] Department of Computer Engineering, Islamic Azad University, Maragheh, Iran,
[3] Department of Computer Engineering, Islamic Azad University, Iran.
[4] Corresponding author, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran,

* Corresponding Author J.nourmohammadi92@ms.tabrizu.ac.ir
ORCID: 0000-0002-1928-9081

(First received on 13 April 2018 and final form in 09 May 2018)

Keywords
Plant Science
Proteomics
Big data
Machine Learning

Abstract: Data extensions in plant biology and drastically increasing data volume in this field impose the scientists analyzing data by means of smart computer systems. Since, manually analyzing huge amount of data is cumbersome and even impossible. A comparative study of proteins a wide scale, is the proteomics knowledge. Nowadays, the proteomics analysis is considered as one of the most important methods in genomics and of the gene expression studies. Large amounts of data are big challenges in plant biology. Biological communities either need to create data making compatible with the parallel computing and the data management associated with its infrastructures or are looking for novel analytical patterns to extract information from a large amount of data. Machine learning provides promising analytical and computational solutions for large, heterogeneous, non-structured datasets for large-scale data, especially for the proteomics data. In particular, a conceptual review and applicable methods of machine learning are described by predicting that how machine learning with massive data technology can be an interface to facilitate basic researches and biotechnology plant sciences.

1. Introduction

The term “proteome” is a combination of two words, “protein” and “genome” [1]. Indeed, these types of terms about Genome are confusing. However, an object includes a Genome; it is able to have various transcripts and proteomes which belong to tissues, formation phases and different conditions. Proteomics is a subject about data processing, analysing and mapping of all proteins which are expressed in a tissue or in a particular cell [2]. Although proteomics is a discriminative method, it is usually used for comparing protein profiles which are expressed in various conditions or tissues. In proteomics analyses, 4 major steps are carried out:

- Performing a method for separating and purifying a small amount of proteins among other mixed cellular macromolecules [3].
- Performing a method for obtaining sequential information about each protein (such as sub-sequential protein, combined amino acid, peptide mass spectrum and etc.) [3].
- Accessing sequential proteins databases or DNA [3].
- Communication of DNA information and information related to sequence, structure, and mechanism of proteins [3].

The machine learning methods have been widely used for analysing data in numerous cases of Biology [4]. More specifically, it is used for comparing a sample of different methods of machine learning methods for data of physiological positions which Biomarkers are made by analysing methods of transcriptomics measurements and proteomics provides a suitable method for this purpose by quantitative assessment of proteins.

Proteomics supplies some of the advantages of data processing, which can be used in both biological cell less fluid, like serum, urine and synovial fluid. Depending on the purpose of the study, expressing a gene does not relate to protein level necessarily, nevertheless, the value of this technology depends on the quality of analysing methods used for produced data processing [5]. Metabolic can be utilized for classification of unknown samples and recognition of disease-related genes. There are similar methods existing about proteomics utilization fields and more specifically, analysing the produced data from a subsequent mass spectrum [6].

An effective mechanism of machines is shown by discoveries in large-scale data from aggregated data source, elements encyclopaedia projects (Encode DNA) and organism model encyclopaedia of DNA components in animals. Nevertheless, machine learning is not vastly used for analysing big data in plants [7]. Nowadays it is an appropriate time for botanical researchers to solve existing problems of plants by using big data technology in their initial research [8].

Despite encouraging the potential of machine learning, it is often misused or mistaken by biologists which is chiefly because of their insufficient knowledge of machine learning and case study of the complexity of biological systems.

Hence, rudimentary objectives of this review try to introduce machine learning basic concepts and methods in biology and conceptualization of machine learning mechanism. Also, it will describe existing practical software for analysing proteomics data [9].

The further subject which will be reviewed in the paper is data mining in proteomics data [10]. Data mining is vastly used in various areas [11-13]. Currently, data mining allows analysing several challenges of companies and organizations. Recently, biological majors like genomics, proteomics, functional genomics and related research had an astonishing growth. Biological data mining is one of the most important fields of bioinformatics. The reasons for using data mining in biology are followed by:

- Levelling, indexing and homogeneous analyses on several Nucleotides sequences
- Discovering structure patterns and analysing Genetics Networks and proteins
- Analysing representation genetics data tools

2. Types of Machine Learning Algorithms

Three major types of machine learning algorithms are as followed:

2.1 Supervised learning

Two types of variables are involved in this type of algorithms which undertakes the most significant role of machine learning (from quantity aspect of algorithms). The first type is called independent variables, there is one or there are some variables which are supposed to predict another variable based their values, for instance, customer age, education, income and marital status are independent variables to predict purchasing goods by a vendee. The second type of variables is dependent or target or output variables which are supposed to be predicted based on these algorithms. For this reason, there should be a function which receives inputs (independent variables) to produce the acceptable outputs (dependent or target variables). Discovering process of this function which is in fact, discovering relationship between dependent and independent variables, is called “Training Process”, which is applied on available data (data which both dependent and independent variables are clear for example earlier purchases of clients of a shop) and continues until reaching adequate accuracy. The samples of these algorithms are: Regression, decision trees, random forest, K nearest neighbour, logistics regression [14].

2.2 Unsupervised learning

In this type of algorithm, there is no target and output algorithms are unknown. The best example which can be made for this type of algorithm, is clustering population, for instance, population can be divided automatically into similar and homogeneous groups by having their personal information and purchases. The Apriori and K-means are of this type [12].

2.3 Reinforcement learning

The third type of algorithms which could be categorized as unsupervised algorithms, are called “Reinforcement Learning". In this type of algorithm a machine (in fact its controller program) is trained for deciding a unique decision and it does based on its current status (a group of available variables), permitted reactions (forward and backward movements) which can be random in incipient levels and for each reaction or behaviour which is returned, system gives a feedback or score to machine and it perceives whether it decided well or not based on feedbacks and it will repeat the same action or try another behaviour in the next similar situations [12].

Due to the affiliation of current state and behavior on prior state and behaviors, Markov decision process can be one of the samples of this group of algorithms, the neural network algorithms can be as well. Purpose of “Reinforcement” in

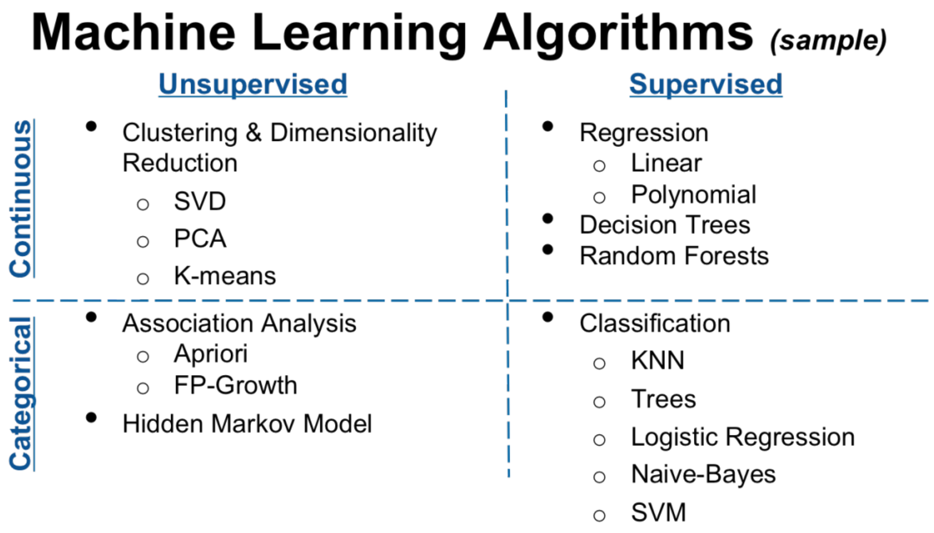


Figure 1. *Classic classification of machine learning algorithms*

naming these algorithms refers to feedback level which improves the mechanism of program and algorithm. A sample of classic classification of machine learning algorithms which are based on supervised and unsupervised and categorical and continuous variables can be seen below (Figure 1).

3. Data Mining

Nowadays as technology is advanced, especially information and communication technology, a large amount of data is generated by communication and information networks and utilizing these data is one of the necessities of businesses, even in small-scale [15]. Different business managers know that collecting information and data from customers and contacts of a company, is one of the vital factors for growing and developing their companies.

However, this is merely half the path, if only data are collected and remain idle, the main objective of data collecting is not practically fulfilled. An important step which has to be taken after collecting information and data is extracting knowledge from data and making collected data more tangible, to deduce practical principles and outcomes.

Collection of tools which can assist companies, institutions and even individuals, in order to reach practical and efficient concepts from volume of data and information, are being studied as Data Mining in a branch of computer science. In fact, data mining is collection of practical problems which are defined in knowledge extraction domain of accessible massive data volumes, also throughout time; methods for solutions are proposed by computer science, mathematics and statistics researchers.

Should Google be able to guess a user's occupation based on his prior activities with a high probability, undoubtedly it is going to show related links to that person's expertise in higher ranks. As another example, existing recommender systems on different websites, take social networks and online shopping stores an instance, for example Amazon, it possesses one of the most powerful recommender systems, having bought a digital camera by customers, they usually order memory cards or tripods as well, this is a pattern that has been gradually discovered by Amazon recommender system, and it provides suggestions to users according to the patterns which are oriented with prerequisites of this pattern. The YouTube or Facebook social network recommender system has almost the same performance and follows the same mechanism [16]. All of these applications which we face with every day are done by analyzing the collected data in the past and they are successful samples of utilizing data mining in daily life. Certainly, these cases are merely examples which can be added easily to other examples.

3.1 Applications of Data Mining

Applications which exist for Data Mining is extensively vast and there is a possibility of introducing a limited number of them in this paper. For further examples, the applications of data mining can be mentioned like below [17]:

- Management systems, like customer relationship management or CRM
- Security software, like network monitoring software and anti-viruses
- Banking systems, credit allocation to customers and their classification

- Economical, like price predicting of one or several stocks or indexes
- Planning and locating, like inner arrangement of malls or urban facility allocations
- Medical science, like predicting the probability of risks of a special surgery operation
- Social and political science, like predicting or analysing the election results

Various issues and applications which eventually lead to data mining are categorized in some major groups. Furthermore, besides declaring and introducing these groups, their applications have been explained by several examples.

Classification: One of the most important capabilities which humans learn and also the most significant part of human knowledge which exactly refers to it. A huge part of human knowledge (both generally and academically) can be modelled as classification. For example, a doctor classifies a patient and puts him into one of the classes and sorts which he has met before, after examining by observing some conditions and doing some measurements,; like “Healthy people group”, “Group of people who caught cold” or “Group of people exposed to seasonal allergy”. Then the doctor writes a prescription related to the class according to his knowledge with some considerations [18]. The data classification method is shown in figure 2.

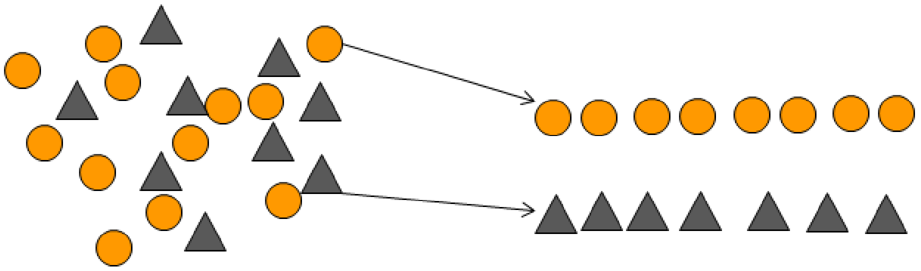


Figure 2. A view of classification

This process can be implemented and simulated by using machine learning methods and special supervised learning methods. For example, if a recommender system in an online shopping store recommends a user buying a novel, one of the possible methods is classifying customers into groups and recommending a special suggestion for each class. It means that if Amazon suggests you to buy a book, from recommender system’s view, you are classified as novel reader people.

Clustering: It is another type of problems which have a close connection with classification. In these issues and applications, no one has solved the problem before and supervisor is unidentified. For example, in different psychological theories, people are divided by different types of personality traits but definitely none of them are neither wrong nor right; they are merely a solution for analysing an analysis. If a computer has to appear in a psychologist’s role, People can be clustered into 5 clusters by a clustering algorithm and by having personality information of several people which are acquired by some examinations. People in a cluster are more similar to each other and at the same time, among two clusters, there is the least similarity (maximum possible difference) [19]. To perform this operation, which is a kind of unsupervised learning, several algorithms have been proposed, such as k-means [19].

Regression: Another family of issues in which, unlike clustering and classification issues, the goal of solving problems is reaching a mathematical equation to describe a phenomenon. For example, the relationship between the time of visiting a site, location, age, used email service, and the amount of order of a user. As another example, the prediction of time series can be referred which can be solved as a special case of regression [20].

3.2 Data mining software and tools

Due to the significance of data mining in the world of large and professional businesses, several tools and software for this purpose have been designed and developed. Some of these tools are free and open source, and others are provided as commercial software packages. Amongst them, a number of things that are mostly used are as follows:

R: Programming language and software package R is one of the most important and effective tools in the fields of statistical deduction and analysis and doing a variety of calculations. The programming language R has a lot of possibilities for performing data mining operations and implementing algorithms related to data mining. This package is completely free and open source [21].

Another important and applicable data mining tool is Microsoft Excel. Performing data mining operations can be provided by default or sometimes adding some commercial extensions in Microsoft Excel [22].

RapidMiner: The RapidMiner software is specialized data mining software that has provided performing various data mining operations, machine learning, text processing, forecasting, and financial analysis tools. Older versions of this software have been published open-source, but the new versions have been commercially available [23].

Weka: The Weka package is also open source software implemented in Java, and a team based at the University of New Zealand Waikato is responsible for the development and maintenance of this software package. This software package has been implemented exclusively for machine learning operations, which of course, has many applications in the field of data mining. This package has been distributed free and open source [24].

MATLAB: MATLAB software and application programming language (MATLAB), as a very popular application, has many features for a variety of majors, including statistical analysis, machine learning, fuzzy systems, artificial neural networks, Modelling, Optimization and Prediction, all of which have vast applications in data mining. Along with the capabilities of the MATLAB, the new MATLAB programming language can also be implemented with new algorithms. The MATLAB main core is distributed commercially. However, some libraries and free toolboxes can also be used for data mining operations by various research and academic groups [25].

IBM SPSS: The IBM SPSS package for statistical analysis (IBM SPSS Statistics), and data mining and modelling (IBM SPSS Modeller) provides a powerful set of tools for various data mining operations. These software applications are commercially available [26].

In general, to understand machine learning applications in proteomics data and its processing, Figure 1 provides a review of these steps. It comes to view that first step is about feeding input in mass spectrometry which needs the most appropriate samples. Mass spectrometry identifies each protein and also quantification is used for measuring proteins.

Second step tries to analyse data and process it to the next step for doing quantification on each protein. The most important part which is more related to the present research is machine learning part. Machine learning methods are used for data analysing, classification and clustering purposes. After applying machine learning methods on data in that process, it is necessary to do post machine learning analyses. This step is done because the proteins should be identified eventually.

4. Databases and Tools

4.1 Plant proteomics

Genomic approaches are completed by studying proteins, they extract interactions among proteins. Having done these extractions, applications of mass spectrometry are needed. Mass spectrometry has caused many developments over the past decade. It does some operations on proteins such as: Identification, detection and analysis which needs proteomics data and tools.

Table 1. *Available databases for plant proteomics*

Name	Web Address	University
PRIDE [27]	http://www.ebi.ac.uk/pride/archive/	The European Bioinformatics Institute, (Cambridge, UK)
PeptideAtlas [28]	http://www.peptideatlas.org/	The Institute for Systems Biology (Seattle, USA)
The Global Proteome Machine and Database [29]	http://www.thegpm.org/	The Global Proteome Machine Organization (Us)
MassIVE [30]	http://massive.ucsd.edu/	The Center for Computational Mass Spectrometry (University of California, San Diego)

Recently, the mass spectrometry is faced with a big challenge that is a large volume of data. Some methods have been proposed for feature extraction on data and also big data approaches are utilized for plant mass spectrometry analysis. In the following there are some databases related to plant proteomics which have been described in Table 1.

Table 2 shows some tools which are used in plant proteomics.

Table 2: *Available tools for plant proteomics*

Name	Web Address	University
The Plant Proteomics Database (PPDB) [31]	http://ppdb.tc.cornell.edu/	Cornell University
1001 Proteomes [32]	http://1001proteomes.masc-proteomics.org/	INRA, France
Pep2pro Database [33]	http://fgcz-pep2pro.uzh.ch/	ETH Zurich

4.2 Plants Genomic analysis

In order to understand the functions and the structures of plant genomes, a large scale of datasets are gathered and some tools have been provided for feature extraction and for doing some analyses on these data sets. These datasets and tools are indicated in Table 3 and Table 4 respectively.

Table 3. *Available databases for plant genomic*

Name	Web Address	University
the National Centre for Biotechnology Information (NCBI) [34]	http://www.ncbi.nlm.nih.gov	National Center for Biotechnology Information, U.S.
EMBL-EBI [35]	http://www.ebi.ac.uk	the University of Cambridge
the DNA Databank of Japan (DDBJ) [36]	http://www.ddbj.nig.ac.jp	Research Organization of Information and Systems National Institute of Genetics DNA Data Bank of Japan (DDBJ)
TAIR [37]	https://www.arabidopsis.org/	The Ohio State University
MaizeGDB [38]	http://www.maizegdb.org/	Iowa State University

Table 4. *Available tools for plant genomic*

Name	Web Address	University
SOL Genomics Network (SOL-GN) [39]	http://solgenomics.net/	Boyce Thompson Institute for Plant Research, USA
Gramene [40]	http://www.gramene.org/	Oregon State University
GreenPhyl DB [41]	http://www.greenphyl.org/cgi-bin/index.cgi	Generation Challenge Programme
Phytozome [42]	http://phytozome.jgi.doe.gov/pz/portal.html	University of California
PLAZA [43]	http://plaza.psb.ugent.be/	Ghent University

5. Discussion

Several methods have been proposed to highlight significant peptides/Mass spectrometry peaks. Figure .3 shows an overview of the topics covered in this review, including the general work flow required and the major considerations that are necessary before beginning an investigation

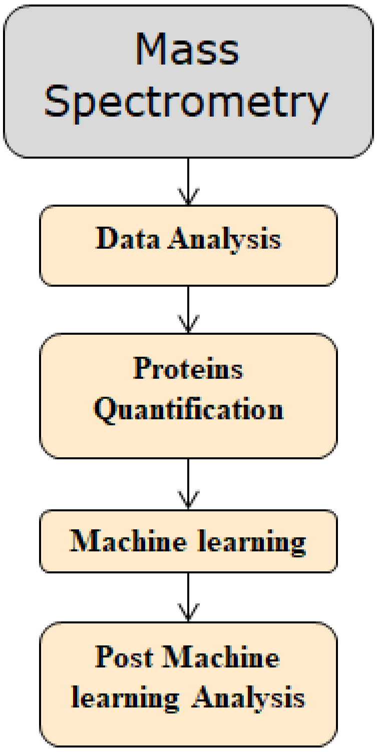


Figure 3. *Machine learning applications in proteomics data and its processing*

combining mass spectrometry and machine learning. A survey of articles involving the combination of mass spectrometry and machine learning will be followed by a brief discussion of post-machine learning analysis, including literature mining and pathway analysis. Machine learning methods have been applied to quantify all proteins; they are used for protein identification too. In the following, some methods are described to discern the problems.

In [44] the authors focused on feature selection methods and classification in proteomics data. They used support machine vector (SVM) and K-Nearest Neighbor (KNN) as a classifier. Their proposed method has more stable feature than REF. They have a name for the proposed method, RELIEF, which is combined with SVM and KNN separately. RELIEF is used for feature extraction while SVM and KNN are used for classification. Table 5 shows average sensitivity, accuracy and specificity of SVM with REF and RELIEF. In [45] a classifier is proposed that is composed of a group of logistic regression to classified cancer samples. Their proposed method has three step-protocols which analyzes mass spectrometry data. The proposed algorithm was applied on a prostate cancer dataset. Authors in [46] used wavelet transforms application in smoothing, peak detection and quality control of mass spectrometry data processing.

Table 5. *Average sensitivity, accuracy and specificity of SVM with REF and RELIEF [44]*

Method Name	sensitivity	accuracy	specificity
RELIEF+ SVM	0.85	0.95	0.90
RFE+SVM	0.93	0.97	0.95
RELIEF union RFE+SVM	0.93	1.0	0.96

They also proposed a new discrete wavelet transform (DWT) which is suitable for smoothing.

In [47] tried to handle the amount of computational effort which is spent on the protein identification. They share their experience in handling of large scale of mass spectrometry data and endeavor to use collateral multi-processing architecture.

Facing with small datasets and based on their high dimensionality of features, the most important challenge in mass spectrometry is data analyzing. The author in [48] proposed a method that applied a Distance Metric Learning for classifying proteomics data. They also used manifold learning for feature reduction.

Another classifier for mass spectrometry data has been proposed by Pascal et.al in [49] for early detection. They used inversion classification which includes an inversion problem with a combined Bayesian method. It is a hierarchical model and presents suitable results.

6. Conclusion

The most interesting and exciting part of this paper is this part which is about future of big data and machine learning in plant science. As is understandable, plant science grows so rapidly and there are some technologies like “Omics” technologies which have a great potential of discoveries in plant science. Future genomics projects will be included in big data in plant science and according to what we have learnt; machine learning has been successfully applied to human biology like disease detection. We also can use it in plant science which means proteomics data that is used is so less. Another future work I have been studying is about spot detection in proteins. Plant specialists spend a great amount of time on checking and detecting the spots which have been extracted and identified by a “Pd-quest” software, this work requires a lot of time and it wastes a lot of time also it has low accuracy. So machine learning especially deep learning can be used to discover a way to make this software more accurate so it can detect spots accurately without help of a specialist in other words, it means the software can be a specialist..

References

[1] A. Benso, S. Di Carlo, H. Ur Rehman, G. Politano, A. Savino, and P. Suravajhala, "A combined approach for genome wide protein function annotation/prediction," *Proteome science,* vol. 11, no. 1, p. S1, 2013.

[2] K. W. Earley *et al.*, "Gateway□compatible vectors for plant functional genomics and proteomics," *The Plant Journal,* vol. 45, no. 4, pp. 616-629, 2006.

[3] P. A. Rudnick *et al.*, "Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses," *Molecular & Cellular Proteomics,* vol. 9, no. 2, pp. 225-241, 2010.

[4] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS computational biology,* vol. 3, no. 6, p. e116, 2007.

[5] A. Pandey and M. Mann, "Proteomics to study genes and genomes," *Nature,* vol. 405, no. 6788, pp. 837-846, 2000.

[6] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *science,* vol. 312, no. 5771, pp. 212-217, 2006.

[7] R. Opgen-Rhein and K. Strimmer, "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data," *BMC systems biology,* vol. 1, no. 1, p. 37, 2007.

[8] V. Marx, "Biology: The big challenges of big data," *Nature,* vol. 498, no. 7453, pp. 255-260, 2013.

[9] L. N. Mueller, M.-Y. Brusniak, D. Mani, and R. Aebersold, "An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data," *Journal of proteome research,* vol. 7, no. 01, pp. 51-61, 2008.

[10] M. Wilkins, "Proteomics data mining," *Expert review of proteomics,* vol. 6, no. 6, pp. 599-603, 2009.

[11] H. Lu, R. Setiono, and H. Liu, "Neurorule: A connectionist approach to data mining," *arXiv preprint arXiv:1701.01358,* 2017.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[13] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering,* vol. 26, no. 1, pp. 97-107, 2014.

[14] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315-3323.

[15] A. El Azab, M. A. Mahmood, and A. El-Aziz, "Effectiveness of Web Usage Mining Techniques in Business Application," in *Web Usage Mining Techniques and Applications Across Industries*: IGI Global, 2017, pp. 324-350.

[16] Y.-D. Seo, Y.-G. Kim, E. Lee, and D.-K. Baik, "Personalized recommender system based on friendship strength in social network services," *Expert Systems with Applications,* vol. 69, pp. 135-148, 2017.

[17] N. Gandhi and L. J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture," in *Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on*, 2016, pp. 1-6: IEEE.

[18] S. Haug, A. Michaels, P. Biber, and J. Ostermann, "Plant classification system for crop/weed discrimination without segmentation," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp. 1142-1149: IEEE.

[19] F. Cannarile *et al.*, "An unsupervised clustering method for assessing the degradation state of cutting tools used in the packaging industry," in *on Proceedings of European Safety and Relaibility Conference, ESREL*, 2017.

[20] H. Kaneko and K. Funatsu, "Adaptive soft sensor based on online support vector regression and Bayesian ensemble learning for various states in chemical plants," *Chemometrics and Intelligent Laboratory Systems,* vol. 137, pp. 57-66, 2014.

[21] P. W. Wilson, "FEAR: A software package for frontier efficiency analysis with R," *Socio-economic planning sciences,* vol. 42, no. 4, pp. 247-254, 2008.

[22] A. C. Burns and R. F. Bush, *Basic marketing research using Microsoft Excel data analysis*. Prentice Hall Press, 2007.

[23] J. Han, J. C. Rodriguez, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," in *Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on*, 2008, vol. 3, pp. 96-99: IEEE.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[25] W. Menke, *Geophysical data analysis: discrete inverse theory: MATLAB edition*. Academic press, 2012.

[26] A. Bryman and D. Cramer, *Quantitative data analysis with IBM SPSS 17, 18 and 19*. Routledge, 2011.

[27] N. del Toro *et al.*, "PRIDE Proteomes: a condensed view of the plethora of public proteomics data available in the PRIDE repository," *DILS 2014,* p. 21, 2014.

[28] U. Kusebauch, E. W. Deutsch, D. S. Campbell, Z. Sun, T. Farrah, and R. L. Moritz, "Using PeptideAtlas, SRMAtlas, and PASSEL: comprehensive resources for discovery and targeted proteomics," *Current protocols in bioinformatics,* pp. 13.25. 1-13.25. 28, 2014.

[29] D. Fenyö and R. C. Beavis, "The GPMDB REST interface," *Bioinformatics,* vol. 31, no. 12, pp. 2056-2058, 2015.

[30] H. Sakai *et al.*, "Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics," *Plant and Cell Physiology,* vol. 54, no. 2, pp. e6-e6, 2013.

[31] Q. Sun, B. Zybailov, W. Majeran, G. Friso, P. D. B. Olinares, and K. J. van Wijk, "PPDB, the plant proteomics database at Cornell," *Nucleic acids research,* vol. 37, no. suppl_1, pp. D969-D974, 2008.

[32] H. J. Joshi *et al.*, "1001 Proteomes: a functional proteomics portal for the analysis of Arabidopsis thaliana accessions," *Bioinformatics,* vol. 28, no. 10, pp. 1303-1306, 2012.

[33] M. Hirsch-Hoffmann, W. Gruissem, and K. Baerenfaller, "pep2pro: the high-throughput proteomics data processing, analysis, and visualization tool," *Frontiers in plant science,* vol. 3, 2012.

[34] D. L. Wheeler *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D13-D21, 2007.

[35] W. Li *et al.*, "The EMBL-EBI bioinformatics web and programmatic tools framework," *Nucleic acids research,* vol. 43, no. W1, pp. W580-W584, 2015.

[36] Y. Tateno *et al.*, "DNA Data Bank of Japan (DDBJ) for genome scale research in life science," *Nucleic acids research,* vol. 30, no. 1, pp. 27-30, 2002.

[37] R. L. Poole, "The TAIR database," *Plant Bioinformatics: Methods and Protocols,* pp. 179-212, 2007.

[38] C. M. Andorf *et al.*, "MaizeGDB update: new tools, data and interface for the maize model organism database," *Nucleic acids research,* vol. 44, no. D1, pp. D1195-D1201, 2015.

[39] L. A. Mueller *et al.*, "The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond," *Plant physiology,* vol. 138, no. 3, pp. 1310-1317, 2005.

[40] M. K. Monaco *et al.*, "Gramene 2013: comparative plant genomics resources," *Nucleic acids research,* vol. 42, no. D1, pp. D1193-D1199, 2014.

[41] M. G. Conte, S. Gaillard, N. Lanau, M. Rouard, and C. Périn, "GreenPhylDB: a database for plant comparative genomics," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D991-D998, 2007.

[42] D. M. Goodstein *et al.*, "Phytozome: a comparative platform for green plant genomics," *Nucleic acids research,* vol. 40, no. D1, pp. D1178-D1186, 2011.

[43] M. Van Bel *et al.*, "Dissecting plant genomes with the PLAZA comparative genomics platform," *Plant physiology,* p. pp. 111.189514, 2011.

[44] E. Marchiori, N. H. Heegaard, M. West-Nielsen, and C. R. Jimenez, "Feature selection for classification with proteomic data of mixed quality," in *Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB'05. Proceedings of the 2005 IEEE Symposium on*, 2005, pp. 1-7: IEEE.

[45] Z. Liu and S. Lin, "Classification Using Mass Spectrometry Proteomic Data with Kernel-Based Algorithms," *Engineering Letters,* vol. 13, no. 4, 2006.

[46] P. Du, S. M. Lin, W. A. Kibbe, and H. Wang, "Application of wavelet transform to the ms-based proteomics data preprocessing," in *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, 2007, pp. 680-686: IEEE.

[47] H. Grover and V. Gopalakrishnan, "Efficient processing of models for large-scale shotgun proteomics data," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, 2012, pp. 591-596: IEEE.

[48] Q. Liu, M. Qiao, and A. H. Sung, "Distance metric learning and support vector machines for classification of mass spectrometry proteomics data," *International Journal of Knowledge Engineering and Soft Data Paradigms,* vol. 1, no. 3, pp. 216-226, 2009.

[49] P. Szacherski *et al.*, "Classification of proteomic ms data as bayesian solution of an inverse problem," *IEEE Access,* vol. 2, pp. 1248-1262, 2014.