INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print), 2222-5234 (Online)
http://www.innspub.net
Vol. 20, No. 5, p. 45-52, 2022

RESEARCH PAPER OPEN ACCESS

Semi-supervised ordered weighted average fuzzy-rough nearest neighbour classifier for cancer pattern classification from gene expression data

Ansuman Kumar[*], **Anindya Halder**

Department of Computer Application, North-Eastern Hill University, Tura Campus, Meghalaya - 794002, India

Key words: Classification, Fuzzy set, Rough set, Ordered weighted average fuzzy-rough, Gene expression cancer data.

http://dx.doi.org/10.12692/ijb/20.5.45-52 Article published on May 24, 2022

Abstract

Classification of cancer patterns from gene expression data is a difficult task in computational biology and artificial intelligence due to the sufficient number of training samples is often difficult, expensive, and hard to gather. Although, the classification results obtained by the conventional classifiers trained with insufficient training samples are generally low. However, unlabeled samples are relatively low-cost and easy to gather, whereas conventional classifiers do not utilize these unlabeled samples to train the model. In this context, a self-training-based model semi-supervised ordered weighted average fuzzy-rough nearest neighbour classifier for cancer pattern classification from gene expression data is proposed. The experiments are carried out on eight publicly available real-life gene expression cancer datasets. The performance of the proposed method is compared with four other methods (two supervised and two semi-supervised) in terms of percentage accuracy, precision, recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa. The dominance of the proposed method is justified by the experimental results.

[*] **Corresponding Author:** Ansuman Kumar ✉ ansuman.kumar@gmail.com

Introduction

Cancer is an unfavourable health issue across the globe. There were approximately 18.1 million new cancer patients and 9.9 million cancer-related deaths worldwide reported in the year 2020, according to the Global cancer statistics**.** Therefore, detection and diagnosis of cancer in an early stage is an important field of research for computational biologists. Conventional techniques for cancer classification depend on the clinical tests and the morphological exhibition of the tumor. These techniques are expensive and time-consuming. The latest development of microarray technology (Stekel, 2003) has enabled scientists to specify ten thousand genes in a single experiment in order to produce a comparatively low-cost diagnosis and prediction of cancer at an early stage. Several machine learning techniques have been applied to classify cancer from microarray gene expression data using supervised learning (i.e., classification) (Dettling *et al.*, 2003) and unsupervised learning (i.e., clustering) (Jiang *et al.,* 2004). However, limited contributions have been made using semi-supervised learning (Priscilla *et al.,* 2013; Halder *et al.*, 2014).

Conversion classifiers need a sufficient number of labeled patterns to train the classifier to achieve the desired accuracy without using unlabeled patterns during the training stage. However, labeled patterns are very costly and hard to collect, whereas unlabeled patterns are relatively low cost and easy to collect. Generally, the number of samples present in gene expression data is very low in comparison to the number of genes available in the dataset. Gene expression datasets are usually vague, indiscernible, and overlapping in nature (Du *et al.,* 2014).

In this context, it is crucial to build a classifier based on a fuzzy-rough set to handle the overlapping, vague and indiscernible subtype classes of gene expression datasets and a semi-supervised learning method that should be useful when the number of labeled patterns is limited. Therefore, in this article, we propose a 'self-training' based semi-supervised ordered weighted average fuzzy-rough nearest neighbour classifier for cancer pattern classification from gene expression data. Semi-supervised learning finds 'high confidence' patterns from unlabeled patterns and adds them to the limited training set to improve classification accuracy.

Material and methods

The proposed semi-supervised ordered weighted average fuzzy-rough nearest neighbour (SS-OWAFRNN) is an amalgamation of fuzzy set and rough set theory; thus,a brief outline of those is provided below:

Fuzzy set theory

L. A. Zadeh developed the fuzzy set (Zadeh, 1965) theory in the year 1965. It is an extension of crisp sets to handle vague and imprecise data. Fuzzy set *A* uses mapping from the universe *X* to the interval [0, 1]. The value of *A*(x) is called the membership degree of x in *A*.

Rough set theory

Rough set theory was introduced by Z. Pawlak (Pawlak, 1982) in the early 1980s. It can handle uncertainty, indiscernibility and incompleteness in the datasets. The rough set theory begins with the idea of an approximation space, which is a pair <*X, R*>, where *X* is the non-empty universe of discourse and *R* is an equivalence relation defined on *X*, where *R* satisfies the reflexive, symmetric and transitive property. The lower and upper approximations for each subset *A* of *X* are defined as follows; the lower approximation is the union of all the equivalence classes which are fully included inside class *A*, and the upper approximation is the union of equivalence classes that have a non-empty intersection with the class *A*.

Fuzzy-rough set theory

Fuzzy set theory can handle vague information, whereas rough set theory can handle incomplete information. Hybridization of these two concepts yields the idea of the fuzzy-rough set, which is the pair of lower and upper approximations of a fuzzy set *A* in a universe *X* on which a fuzzy relation *R* is defined. The fuzzy-rough lower and upper approximations of A are defined as follows (Cornelis *et al.,* 2010):

$$(R\downarrow A)(x)=\inf_{y\in X} I(R(x,y),A(y)) \quad (1)$$

$$(R\uparrow A)(x)=\sup_{y\in X} T(R(x,y),A(y)) \quad (2)$$

Where I is the Lukasiewicz implicator, T is the Lukasiewicz t-norms and $R(x,y)$ is the valued similarity of patterns x and y, inf is infimum and sup represents supremum (Cornelis *et al.*,2010).

Ordered weighted averaging fuzzy-rough sets theory

Fuzzy-rough set theory used Lukasiewicz implicator and *t*-norm to compute upper and lower approximations based on only one instance by using inf and sup operators. Therefore, Ordered Weighted averaging (OWA) operators modify the strict inf and sup operators with more flexible operators to obtain lower and upper approximations that would be more robust in the presence of noise (Cornelis *et al.*, 2010). The lower and upper approximations of instance y can be defined as follows:

$$(R\downarrow_{OWA} A)=OWA_{\min} I(R(x,y),A(y)) \quad (3)$$

$$(R\uparrow_{OWA} A)=OWA_{\max} T(R(x,y),A(y)) \quad (4)$$

Proposed Semi-supervised Ordered Weighted Average Fuzzy-rough Nearest Neighbour Classifier

The proposed method semi-supervised Ordered Weighted Average Fuzzy-rough Nearest Neighbour Classifier (SS-OWAFRNN) comprises semi-supervised learning and testing. In the first stage, semi-supervised learning is adopted to select the 'high-confidence' patterns from the unlabeled patterns and it is used for the training process in the next iteration. This process continues until convergence. In the second stage, each test pattern is assigned to a particular class based on the final enlarged training set (limited numbers of the labeled patterns together with the high-confidence unlabeled patterns). The complete procedure of SS-OWAFRNN method is given below:

Stage-I: Semi-supervised learning

A. Determine the *k*-nearest neighbour (*kNN*) labeled patterns nearest to each of the unlabeled pattern (*u*) based on Euclidean distance.

B. The values of lower and upper approximations of unlabeled pattern (*u*) for belonging to each class *C* is computed using Equation (5) and (6) respectively (Jensen *et al.*, 2008):

$$(R\downarrow_{OWA} C)=OWA_{\min} I(R(u,y),C(y)) \quad (5)$$

$$(R\uparrow_{OWA} C)=OWA_{\max} T(R(u,y),C(y)) \quad (6)$$

Where $R(u,y)$ is calculated as follows:

$$R(u,y)=\frac{\sum_{y\in kNN}(\|u-y\|)^{\frac{2}{m-1}}}{(\|u-y\|)^{\frac{2}{m-1}}}; \quad (7)$$

where $\|u-y\|$ is the distance of the unlabeled pattern (u) from the labeled pattern $y\in kNN$ (*k*-nearest neighbour labeled pattern of unlabeled pattern u) and m $(1<m<\infty)$ is the fuzzifier. $C(y)$ is calculated as:

$$C(y)=\begin{cases}1, & if \quad y\in C;\\ 0, & Otherwise.\end{cases} \quad (8)$$

$OWA_{\min}$ and $OWA_{\max}$ weights are calculated as:

$$OWA_{\min}=\left\langle \frac{2}{p(p+1)},\frac{4}{p(p+1)},\ldots\ldots,\frac{2p}{p(p+1)}\right\rangle \quad (9)$$

$$OWA_{\max} = \left\langle \frac{2p}{p(p+1)}, \frac{2(p-1)}{p(p+1)}, \ldots\ldots, \frac{2}{p(p+1)} \right\rangle \qquad (10)$$

where $p = |kNN|$.

C. Unlabeled samples having high-confidence values are picked as:

Let avj_{ij} be the average value of lower and upper approximations of the unlabeled sample u_i for belonging to a class C_j and $\max_h$ be the highest average value of lower and upper approximations of the unlabeled pattern u_i for belonging to a class C_h.

The ratio $\left(\frac{avg_{ij}}{\max_h}\right)(\forall j,\ j \neq h)$ acts the degree of similarity of an unlabeled sample u_i for belonging to class C_j and the highest belonging class C_h . The value of the ratio lies between 0 and 1. Higher the value of the ratio, more is the similarity of the unlabeled sample with two classes C_j and C_h ; thus, less is the confidence of that unlabeled sample for belonging to any class. Therefore, if the ratio value of avj_{ij} $(\forall j,\ j \neq h)$ and $\max_h$ is less than threshold value (close to 0) then the corresponding sample is considered as high confidence pattern u_i (belonging to class C_h) and added to the training set for the next iteration. Otherwise, u_i is not added to the next iteration.

Stage-II: Testing

Stage-I (semi-supervised learning) is once converged, then the test patterns are tested to assign the class labels based on the enlarged set of labeled patterns.

A. Compute the *k*-nearest neighbour (*kNN*) labeled patterns nearest to each of the test pattern (t) based on Euclidean distance.

B. The values of lower and upper approximations of a test pattern (t) for belonging to each class *C* is calculated (similar to Equation (6) and (7)) respectively as follows (Jensen *et al.*, 2008):

$$(R\downarrow_{OWA} C) = OWA_{\min} I(R(t,y), C(y)) \qquad (11)$$

$$(R\uparrow_{OWA} C) = OWA_{\max} T(R(t,y), C(y)) \qquad (12)$$

where $R(t,y)$ is computed as follows:

$$R(t,y) = \frac{\sum_{y \in kNN} (\|t-y\|)^{\frac{2}{m-1}}}{(\|t-y\|)^{\frac{2}{m-1}}}; \qquad (13)$$

where $\|t-y\|$ is the distance of the test pattern (t) from the labeled pattern $y \in kNN$ (*k*-nearest neighbour labeled pattern of test pattern (t) and m $(1 < m < \infty)$ is the fuzzifier. $C(y)$ is computed using Equation (8). $OWA_{\min}$ and $OWA_{\max}$ are computed using Equation (9) and (10), respectively.

C. Test pattern (t) is conferred to a particular class for which the sum of lower and upper approximation value is highest.

$$ClassLabel(t) = \arg\max_{j} \begin{pmatrix} (R\downarrow_{OWA} C_j)(t) + \\ (R\uparrow_{OWA} C_j)(t) \end{pmatrix} ;\forall t \qquad (14)$$

Comparison with other methods

The comparison of the proposed SS-OWAFRNN method with four other methods, namely, fuzzy *k*-nearest neighbour (FKNN) (Keller *et al.,* 1985), ordered weighted average fuzzy-rough nearest neighbour (OWAFRNN) (Cornelis *et al.,* 2010), semi-supervised fuzzy *k*-nearest neighbour (SS-FKNN) (Halder *et al.,* 2014), and semi-supervised fuzzy vaguely quantified rough nearest neighbour (SS-

FVQRNN) (Jensen *et al.*, 2008) is done. Fuzzy *k*-nearest neighbour (FKNN) (Keller *et al.,* 1985) method is a continuation of the *k*-Nearest Neighbour (KNN) classifier. In *KNN* algorithm, equal weightage is provided to all the *k*-nearest neighbours to compute the predicted class of a test sample. FKNN algorithm assigns fuzzy membership of a test sample in each class. That class is considered to be the predicted class (of that test sample) for which the fuzzy-membership is maximum. SS-FKNN (Halder *et al.,* 2014) method is the semi-supervised version of FKNN method, which utilizes the unlabeled samples along with the labeled samples to enhance the classification accuracy of the cancer classification.

In SS-FVQRNN (Jensen *et al.*, 2008) method, the test sample is assigned to a specific class for which the sum of lower and upper approximation value is highest and this method is a semi-supervised version of FVQRNN method.

Performance evaluation measures

In this article, six validity measures are used to evaluate the performance of the proposed method. They are (i) percentage accuracy, (ii) precision, (iii) recall, (iv) macro averaged F_1 measure, (v) micro averaged F_1 measure (Halder *et al.*, 2013) and (vi) kappa (Cohen, 1960).

Experimental setup

The average results of 10 simulation runs of all the methods carried out on eight real-life microarray gene expression datasets are mentioned in this article.

All the methods are implemented in MATLAB and executed on Windows 7 machine with a processor speed 2.40 GHz and main memory 4 GB in this article. Two samples are taken from each class as a training sample and the test set comprises the total samples available (in the datasets), keeping out the training samples.

Description of datasets

We have used eight real-life gene expression cancer datasets, namely, Colon Cancer, Brain tumor, SRBCT, Lymphoma, Prostate Cancer, Ovarian Cancer, Leukemia, Lung Cancer datasets for the experiments. These datasets are publicly available at www.stat.ethz.ch/dettling/bagboost.html (Dettling, 2004) and http://datam.i2r.astar.edu.sg/datasets/krbd/index.html (Kent ridgebio-medicaldataset repository).

A dataset is a group of samples and each sample has gene expression values and class information. A brief outline of the used datasets is given below.

The colon Cancer dataset consists of 40 samples, out of which 22 samples are normal patients and 18 samples are cancerous. Each sample has 2000 genes.

The brain Tumor dataset contains 42 samples separated into 5 classes (viz., medulloblastomas, malignant gliomas, atypical teratoid/rhabdoid tumors, primitive neuroectodermal tumors, human cerebella). The numbers of samples present for these classes are 10, 10, 10, 8 and 4, respectively. There are 5597 genes expression values in each sample.

The small round blue cell tumors (SRBCT) dataset contains 63 samples, out of which 12 samples of neuroblastoma, 20 samples of rhabdomyosarcoma, 8 samples of Burkitt's lymphoma and 23 samples of Ewing's sarcoma. Each sample is described by 2308 genes.

The lymphoma dataset has 62 samples and each sample is described by 4026 genes. There are 3 classes of lymphoma viz., diffuse large B-cell lymphoma, follicular lymphoma and chronic lymphocytic leukemia.

The prostate cancer dataset comprises 102 samples, of which 52 samples are from prostate cancer tissues and 50 samples are from normal patients. Each sample contains expression values for 6033 genes.

The ovarian cancer dataset contains 203 observations, of which 91 observations are normal and 162 observations are cancerous. The number of genes present in each observation is 15154. *The leukemia dataset* contains 72 samples with two classes, namely, lymphoblastic leukemia and myeloid leukemia and each sample is described by 3571 genes.

The lung Cancer dataset has 203 samples of which 139 samples of lung adenocarcinomas, 20 samples of pulmonary carcinoids, 21 samples of squamous cell lung carcinomas, 6 samples of small-cell lung carcinomas and 17 normal lung samples. The expression profile contains 12600 genes. The datasets used for the experiments are summarized in Table 1.

Results and discussion

The average results of 10 simulation run (on a random selection of labeled / training samples) in terms of percentage accuracy, precision, recall, macro F_1, micro F_1 and kappa obtained by all the methods (viz., FKNN, OWAFRNN, SS-FKNN, SS-FVQRNN and SS-OWAFRNN) achieved on eight microarray gene expression datasets are shown in Table 2. The best results are shown in bold font in Table 2. The standard deviations of accuracies of 10 simulations are also shown using ± sign corresponding to each percentage accuracy in Table 2.

Table 1. Summary of eight microarray gene expression datasets used for the experiments.

Datasets	No. of Samples	No. of Genes	Classes
Colon Cancer	62	2000	2
Brain Tumor	42	5597	5
SRBCT	63	2308	4
Lymphoma	62	4026	3
Prostate cancer	102	6033	2
Ovarian cancer	253	15154	2
Leukemia	72	3571	2
Lung Cancer	203	12600	5

From the Table 2, it is observed that the proposed SS-OWAFRNN method performed better in terms of all most all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) compared to the other methods namely, FKNN, OWAFRNN, SS-FKNN and SS-FVQRNN for six datasets (viz., Colon cancer, Brain Tumor, SRBCT, Lymphoma, Prostate cancer, Ovarian cancer and Leukemia). Whereas only one Lung cancer dataset, SS-FVQRNN method achieved better results in terms of all the validity measures compared to the other methods.

Table 2. Summary of the average experimental results (in terms of accuracy, precision, recall, macro F_1, micro F_1 and kappa) of 10 simulations obtained by different methods viz., FKNN, OWAFRNN, SS-FKNN, SS-FVQRNN and SS-OWAFRNN performed on eight microarray gene expression datasets.

Datasets	Methods	Accuracy (%)	Overall Precision	Overall Recall	MacroF_1	Micro F_1	Kappa
Colon cancer	FKNN	80.69 ± 8.28	0.8467	0.8237	0.8029	0.8350	0.6255
	OWAFRNN	90.52 ± 6.20	0.9158	0.9056	0.9001	0.9105	0.8039
	SS-FKNN	85.52 ± 8.45	0.8824	0.8577	0.851	0.8698	0.7119
	SS-FVQRNN	93.45 ± 2.67	0.9364	0.9238	0.9303	0.9349	0.8614
	SS-OWAFRNN	94.83 ± 4.45	0.9468	0.9480	0.9421	0.9424	0.8842
Brain Tumor	FKNN	67.81 ± 7.66	0.6692	0.7901	0.6433	0.7224	0.5812
	OWAFRNN	80.05 ± 4.19	0.7836	0.8387	0.7692	0.8091	0.7407
	SS-FKNN	75.31 ± 9.93	0.7117	0.7934	0.6964	0.7496	0.6773
	SS-FVQRNN	80.21 ± 4.37	0.7893	0.8177	0.7640	0.8028	0.7443

	SS-OWAFRNN	81.82 ± 3.61	0.8083	0.8800	0.7799	0.8426	0.7612
SRBCT	FKNN	71.45 ± 4.37	0.7918	0.7727	0.7140	0.7818	0.6239
	OWAFRNN	81.09 ± 6.37	0.8566	0.8040	0.7947	0.8294	0.7450
	SS-FKNN	78.00 ± 9.37	0.8375	0.7955	0.7634	0.8158	0.7068
	SS-FVQRNN	79.36 ± 4.47	0.8405	0.8051	0.7760	0.8221	0.7223
	SS-OWAFRNN	83.64 ± 3.99	0.8718	0.8167	0.8224	0.8433	0.7762
Lymphoma	FKNN	96.25 ± 1.01	0.9786	0.9218	0.9474	0.9493	0.9202
	OWAFRNN	96.43 ± 1.19	0.9833	0.9261	0.9516	0.9538	0.9241
	SS-FKNN	96.61 ± 0.56	0.9841	0.9293	0.9542	0.9559	0.9275
	SS-FVQRNN	95.18 ± 2.24	0.9461	0.9004	0.9194	0.9226	0.8971
	SS-OWAFRNN	98.21 ± 1.39	0.9917	0.9667	0.9782	0.9790	0.9610
Prostate cancer	FKNN	67.55 ± 10.89	0.6736	0.7444	0.6425	0.7047	0.3471
	OWAFRNN	85.00 ± 5.61	0.8499	0.8738	0.8472	0.8615	0.6997
	SS-FKNN	76.02 ± 9.73	0.7591	0.7932	0.7486	0.7754	0.5186
	SS-FVQRNN	81.96 ± 8.06	0.8211	0.8502	0.8140	0.8350	0.6403
	SS-OWAFRNN	87.76 ± 5.05	0.8754	0.8952	0.8757	0.8852	0.7540
Ovarian cancer	FKNN	87.07 ± 7.50	0.8563	0.8704	0.8525	0.8626	0.7100
	OWAFRNN	89.40 ± 4.92	0.9095	0.8944	0.8887	0.9015	0.7814
	SS-FKNN	88.92 ± 5.55	0.8818	0.8866	0.8798	0.8841	0.7611
	SS-FVQRNN	96.63 ± 4.22	0.9713	0.9630	0.9646	0.9671	0.9298
	SS-OWAFRNN	96.79 ± 5.28	0.9750	0.9588	0.9657	0.9668	0.9314
Leukemia	FKNN	75.59 ± 5.77	0.7879	0.7668	0.7482	0.7772	0.5162
	OWAFRNN	76.91 ± 7.73	0.7395	0.7956	0.7290	0.7637	0.4770
	SS-FKNN	76.03 ± 7.6	0.7987	0.7813	0.7542	0.7899	0.5338
	SS-FVQRNN	73.94 ± 8.55	0.6885	0.7159	0.6865	0.7014	0.3859
	SS-OWAFRNN	80.88 ± 1.87	0.8556	0.8194	0.8054	0.8371	0.6248
Lung cancer	FKNN	61.81 ± 8.43	0.7895	0.6061	0.6070	0.6852	0.4414
	OWAFRNN	65.76 ± 14.41	0.7751	0.6064	0.6133	0.6787	0.4903
	SS-FKNN	71.19 ± 8.61	0.7730	0.6742	0.6576	0.7191	0.5328
	SS-FVQRNN	72.99 ± 6.88	0.8027	0.6505	0.6638	0.7158	0.5526
	SS-OWAFRNN	66.25 ± 4.23	0.7867	0.6066	0.6846	0.6852	0.4607

Conclusion

This article presents a novel ‘self-training’ based semi-supervised Ordered Weighted Average Fuzzy-rough Nearest Neighbour Classifier (SS-OWAFRNN) for cancer sample classification from gene expression datasets. The scarcity of the training samples is handled by the semi-supervised learning technique, whereas overlapping, vague and indiscernibility present in the cancer subtype classes of microarray gene expression datasets are dealt with by the fuzzy and rough set theory in the proposed method. The efficiency of the proposed SS-OWAFRNN method is validated using eight real-life gene expression cancer datasets in terms of six validity measures viz., accuracy, precision, recall, Macro F_1, Micro F_1 and kappa. It is seen from the experimental results that the proposed SS-OWAFRNN method achieved a better result in terms of all most all the validity measures for seven datasets, namely, Colon cancer, Brain Tumor, SRBCT, Lymphoma, Prostate cancer, Ovarian cancer and Leukemia. In contrast, the nearest competitive SS-FVQRNN method performed better in terms of all the validity measures for only one Lung cancer dataset. The promising results achieved from the proposed method motivate us to apply a semi-supervised learning framework to other classifiers. The proposed method may also be validated on other microarray / micro-RNA gene expression cancer datasets in the future.

References

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20(1),** 37–46. https://doi.org/10.1177/001316446002000104

Cornelis C, Verbiest N, Jensen R. 2010. Ordered Weighted Average Based Fuzzy Rough Sets. In: Yu *et al.* Ed. Lecture Notes in Computer Science, Springer, Berlin, Germany **6401,** 78–85. https://doi.org /10.1007/978-3-642-16248-0_16

Dettling M. 2004. Bagboosting for tumor classification with gene expression data. Bioinformatics **20(18),** 583–593. https://doi.org/10.1093/bioinformatics/bth447

Dettling M, Buhlmann P. 2003. Boosting for tumor classification with gene expression data. Bioinformatics **19(9),** 1061–1069. https://doi.org/10.1093/bioinformatics/btf867

Du D, Li K, Li X, Fei M. 2014. A novel forward gene selection algorithm for microarray data. Neurocomputing **133,** 446–458. https://dblp.org/rec/journals/ijon/DuLLF14

Halder A, Ghosh S, Ghosh A. 2013. Aggregation pheromone metaphor for semi-supervised classification, Pattern Recognition **46(8),** 2239–2248. https://doi.org/10.1016/j.patcog.2013.01.002

Halder A, Misra S. 2014. Semi-supervised fuzzy k-NN for cancer classification from microarray gene expression data.In: Proceedings of the 1st International Conference on Automation, Control, Energy and Systems (IEEE Computer Society Press), 1–5. https://doi.org/10.1109/ACES.2014.6808013

Jensen R, Cornelis C. 2008. A new approach to fuzzy-rough nearest neighbour classification. In:Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing, 310–319. https://doi.org/10.1007/978-3-540-88425-5_32

Jiang D, Tang C, Zhang A. 2004. Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering **16 (11),** 1370–1386. https://doi.org/10.1109/TKDE.2004.68

Keller JM, Gray MR, Givens JA. 1985. A fuzzy K-nearest neighbor algorithm, IEEE Transactions on Systems, Man and Cybernetics **15(4),** 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Pawlak Z. 1982. Rough sets. International Journal of Computer and Information Science **11(5),** 341–356. https://doi.org/10.1007/BF01001956

Priscilla R, Swamynathan S. 2013. A semi-supervised hierarchical approach: two-dimensional clustering of microarray gene expression data. Frontiers of Computer Science **7(2),** 204–213. https://doi.org/10.1007/s11704-013-1076-z

Stekel D. 2003.Microarray Bioinformatics. 1st ed., Cambridge, Cambridge University Press, UK. https://doi.org/10.1093/aob/mch083

Zadeh L. 1965. Fuzzy sets. Information and Control **8(3),** 338–353. http://dx.doi.org/10.1016/S0019-9958(65)90241-X