Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

PREDICTING SUBCELLULAR LOCALIZATION OF PROTEINS WITH MULTIPLE SITES USING THRESHOLD ML-KNN

U.SUBHASHINI[*1], P.BHARGAVI[2], S.JYOTHI[3] AND D.M.MAMATHA[4]

*[*1]Research Scholar, Department of Computer Science,*
Sri Padmavati Mahila Visvavidyalayam Tirupati-517 502
[2]Assistant professor, Department of Computer Science,
Sri Padmavati Mahila Visvavidyalayam Tirupati-517 502
[3]Professor, Department of Computer Science,
Sri Padmavati Mahila Visvavidyalayam Tirupati-517 502
[4] Professor, Department of Sericulture,
Sri Padmavati Mahila Visvavidyalayam Tirupati-517 502

ABSTRACT

Predicting the appropriate protein subcellular localization has attracted much attention in the field of bioinformatics for determining the cellular function of proteins. Several traditional biochemical experimental methods have been developed to determine the protein subcellular localization is expensive and time-consuming and during the last decade, many computational based methods have been developed to predict the protein subcellular localization in different organisms. However, most of the methods succeeded to predict proteins in only one subcellular location but there are many proteins which have two or more subcellular locations. To predict subcellular localization of protein with multiple sites we used Threshold-MLKNN (multi-label k-Nearest Neighbours algorithm). Threshold-MLKNN performs much better than MLKNN and produces better prediction accuracy in much lesser time.

KEYWORDS: *Protein subcellular localization, Amino Acid Composition, PseAA, Physicochemical property, ML-KNN, Threshold-MLKNN.*

U.SUBHASHINI*
Research Scholar, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam Tirupati-517 502

***Corresponding Author**

Received on: 10.11.2016
Revised and Accepted on: 23-05-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.3.b278-285

INTRODUCTION

At the turn of this century, researchers are increasingly turning to proteomic studies to understand the biological processes at the cellular level. One of the most important goals in proteomics is protein subcellular localization and that has received a lot of attention recently because it is the key functional attribute of a protein. Information about the subcellular localizationof proteins is important to determine how they interact with each other and with other molecules and knowing the information about protein subcellular localization is important to understand not only the function of proteins but also the organization of the whole cell. Subcellular information can be obtained by conducting various biochemical experiments.[1-2] However, it is time-consuming, expensive, and laborious.[3-4] Nowadays, more and more proteins are found, and it makes such techniques more unpractical. Recent technical advances in large-scale sequencing and genomics procedures have triggered a scientific revolution and resulted in the massive accumulation of proteins whose functions are unknown. Recently many of the genome sequencing projects has called for the development of novel and powerful tools for timely predicting the subcellular location of uncharacterized proteins and numerous computational methods have been made to predict the protein subcellular localization.[5-7] However, many of the methods are succeeded to predict the subcellular locations for a single location. But unfortunately various multi-location proteins are located at more than one location simultaneously. When subcellular localization prediction models are constructed by these methods, multi-location proteins are not included in the training set. Actually, multi-sites proteins have special biological functions, which are helped in the development of new drugs. Thus it is highly needed to use acomputational method which can identify multi sites subcellular localization fast and reliable. In this paper, we used a computational algorithm (Threshold-MLKNN) with different techniques for protein representation and for extracting appropriate features from proteins with which the identification can accurately identify more subcellular locations of proteins and in the meantime bear less unwanted bias. Extracting features from proteins for classification has value in many applications, including subcellular localization.[8-9] Feature extraction is the key process to predict multisite protein subcellular localization and its effective features vector can significantly increase prediction precision.[10]To obtain efficient features from proteins,the protein composition information and amino acid location information must be considered.

Protein Feature Extraction Methods

To computationally analyze the protein sequence data and to successfully use them in ML-KNN and Threshold multi-label k nearest neighbour (Threshold-MLKNN) algorithm, one of the computational challenges is to characterize a protein sequence data by a fixed length feature vector.[11] So it is required to transform the protein sequence data into vectors of numerical values in which the important information of proteins is fully encoded.[12]The three feature extraction methods that are used in this process are:

Amino Acid Composition (AAC)

Using Amino acid composition Protein information can be encapsulated in a vector of 20 dimensions.[13] The Amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 amino acids were calculated as,

$$AAC(i) = \frac{N_i}{N} \qquad i=1,2,..20 \qquad (1)$$

Where N_i is the number of amino acids of type i, N is the protein sequence length.

Pseudo Amino Acid Composition (PseAA)

PseAA composition of a protein is actually a set of distinct numbers that is derived from its amino acid sequence and that is different from the classical AA composition. The pseudo amino acid (PseAA) composition was proposed by for improving the prediction accuracy of protein subcellular localization.[14] PseAA formulated as

$$P = [p_1, p_2, \ldots p_{20}, p_{20+1}, \ldots p_{20+n}], (n < N) \qquad (2)$$

Where P is the protein and $p_1, p_2 \ldots$ are the components. The calculation method is the same as amino acid composition, and the rest n dimension vectors are the location information. Once the composition information and location information are calculated, the feature vector of pseudo amino acid composition model can be represented as follows:

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{n} \delta_k} (1 \le u \le 20) \\ \dfrac{w\delta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{n} \delta_k} (20+1 \le u \le 20+n) \end{cases} \qquad (3)$$

Where w is the weight factor and usually defined as 0.05, f_i is the occurrence frequency of the amino acid residues. δ_k is the k- th tier correlation factor The value of h should be appropriate.μ

Physicochemical Properties Model (PC)

Physichemicalproperties[15] can directly show the activities of a protein. Such properties contain the substantial feature information. Thus, correctly extract the physichemical properties of proteins can effectively improve the predicting performance.[16] The extracting method of physichemical properties resemble the extracting method of amino acid composition, the difference is that we should compute the occurrence frequency in three cases:

$$f_{i,polar} = \frac{n_{polar}}{N} i = 1,2....7 \quad (4)$$

$$f_{i,neutral} = \frac{n_{neutral}}{N} i = 1,2....7 \quad (5)$$

$$f_{i,hydrophobic} = \frac{n_{hydrophobic}}{N} i = 1,2...7 \quad (6)$$

f_i, represents the frequency of amino acid which is characterized by polar, the other two represent the frequencies of neutral and hydrophobic amino acid. N is the protein sequence length. By the fusion of three feature extraction methods we can get the efficient feature vector and this is given as input to the ML-KNN and threshold MLKNN to predict the multiple sites subcellular locations.

Prediction Accuracy Methods

A key step in the subcellular localization prediction is to select an appropriate classification algorithm. In the study of multisite protein subcellular localization, researchers have proposed many multi-label classification algorithms. The threshold MLKNN algorithm is a simple classification algorithm. In spite of its simplicity, it can give competitive performance when compared to ML-KNN.To evaluatethe effectiveness of the fusion feature extraction method two different kinds of classifiers are used to predict protein subcellular localization. The experimental results show that threshold MLKNN achieves superior to existing approaches in prediction accuracy.

ML-KNN

Zhou and Zhang developed the multi-label KNN based on traditional KNN algorithm, which resolve with multi-label learning problems.[17-19]ML-KNNis derived from the traditional k-Nearest Neighbors (k-NN) algorithm. Given an instance x and its associated label set $Z \subseteq z$, suppose k nearest neighboursis considered in the ML-KNN method. Let $N(x)$ denote the set of k Nearest Neighbors of x identified in the training set.Thus, a membership counting vector based on the label sets of k neighbors can be defined as:

$$\vec{A}_x(l) = \sum_{c \in N(x)} \vec{z}_a(l), l \in z \quad (7)$$

Where $\vec{A}_x(l)$ counts the number of neighbours of x belonging to the l-th class. $\vec{A}_x$, be the category vector for x.For each test instance t, ML-KNN identifies its k nearest neighbors *N* (*t*) in the training set. Based on the membership counting vector $\vec{A}_t$, the category vector $\vec{z}_t$ is determined using the following maximum a posterioriprinciple:

$$\vec{z}_t(l) = \arg\max_{b \in \{0,1\}} P(G_b^l \mid H^l_{\vec{A}_t(l)}) \quad (8)$$

Where G_1^l be the event that t has label l, while G_0^l be the event that t has not label l. $H^l_{j \in \{0,..,k\}}$, denotes the event that, among the k nearest neighbors of *t*, there are exactly *j* instances which have label l. Using the Bayesian rule, Eq.(7) can be rewritten as:

$$\vec{z}_t(l) = \arg\max_{b \in \{0,1\}} P(G_b^l) P(H^l_{\vec{A}_t(l)} \mid G_b^l) \quad (9)$$

Threshold ML-KNN

Threshold MLKNN is a modification of ML-KNN classificationalgorithm.[17]The key principle is to first compute the prior probability of a sample class label. Then, ensure the k neighbours of the unknown label according to the k nearest neighbour method. Finally, with the Bayes decision theory determine whether the unknown sample contains the

label.Since some proteins in datasets may occur in two or more locations. According to the concept of ‘‘locative protein’’ , if a protein occurs at two different locations then it will be treated as 2 locative proteins or if it occurs at 3 sites then it will be treated as 3 locative proteins and so forth.[21]Thus, it follows

$$T(loc) = T(seq) + \sum_{m=1}^{\gamma} (m-1)T(m) \qquad (10)$$

where T(loc) is the number of total locative proteins, T(seq) the number of total different protein sequences, T(1) the number of proteins with one location, T(2) the number of proteins with two locations, and so forth; while γ is the number of total subcellular location sites concerned. In it has been shown that, due to noise in the data ML-KNN may perform poorly on a real dataset.[17]A solution to this problem has been proposed is that instead of approximating the Bayes probabilities, a single threshold(τ)can be given to each label l; after selecting a threshold value, label l is assigned to object x when at least τ objects of class l are found in the neighbourhood of x.[18] For a given label l, let us denote by a_l^i number of objects in the training set described by label l, that have exactly i objects with label l assigned in their neighbourhood and b_l^i number of objects in the training set, that have exactly i objects with label l assigned in their neighbourhood and which are not described by label l. For choosing of threshold value, one can traverse through all $\tau \in \{0,\ldots.k\}$and check, for which τ some utility function is the highest. The set of values a_l^i, b_l^i can be measured using the following variable:

1. FN (false negative), number of objects incorrectly classified as not belonging to class l. It can be calculated as $\sum_{i<t} a_l^i$
2. TP (true positives), number of objects correctly classified as belonging to class l. It can be calculated as $\sum_{i\geq t} a_l^i$
3. TN (true negative), number of objects correctly classified as not belonging to class l. It can be calculated as $\sum_{i<t} b_l^i$
4. FP (false positive), number of objects incorrectly classified as belonging to class l. It can be calculated as $\sum_{i\geq t} b_l^i$.

Dataset

We used plant protein dataset, collected from the Swiss-Prot database.[20]Plant protein dataset consists of 978 different protein sequences, which are distributed among 12 plant subcellular locations.

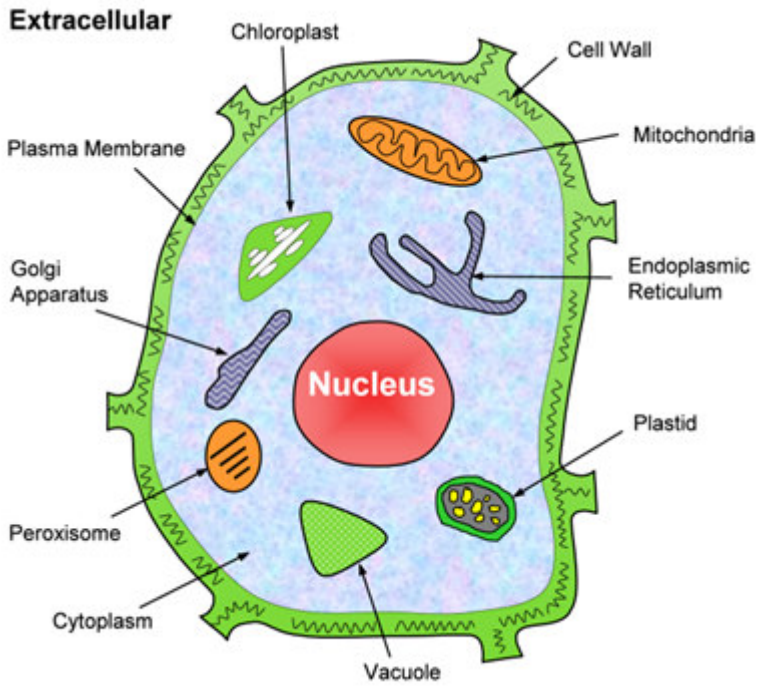

Figure 1
Subcellular Locations of plant proteins

Among the 978 different proteins, 904 belong to one subcellular location, 71 to two locations, and 3 to three locations. Plant proteins subcellular locations are shown in Figure 1 and the numbers of proteins located in different subcellular are shown in Figure2.

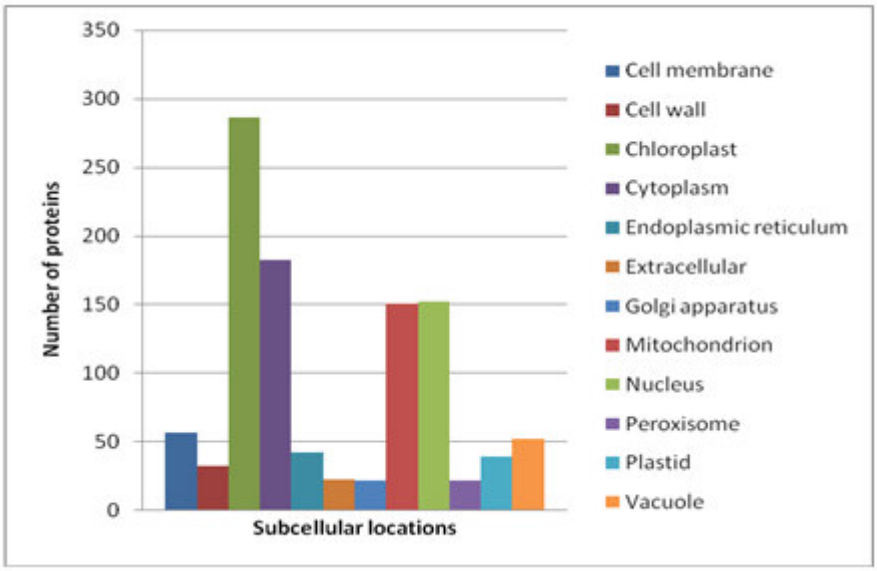

Figure 2
Plant protein benchmark dataset

Experimental Results
In the experiment plant dataset was used and we applied Threshold ML-KNN to predict the multi-site subcellular localization. In the prediction process first, the dataset D containing different protein sequences are distributed according to their subcellular locations.

$$\text{i.e., } D = D_1 \cup D_2 \cup \ldots\ldots\ldots \cup D_n \qquad (11)$$

where D_1 represents the subset for the subcellular location of a cell membrane, D_2 for cell wall, D_3 for chloroplast, and so forth; n is the number of subcellular locations. The sequences were encoded using the combination of Amino Acid Composition, Pseudo Amino Acid Composition and Physicochemical Properties Model. Once the encoding is completed then these sequences were used for training and testing.

Evaluation Criteria
We use the following measures that are used to evaluate the prediction performance of the Threshold ML-KNN.

Hamming loss is defined as**:**

$$hamming\_loss(h) = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{1}{q}\,|\,h(x_i)\Delta L_i\,| \qquad (12)$$

Where m_t is the number of samples and q represents labels in the training set. $h(x_i)$ Predicted label, L_i actual label, and Δ represent the symmetric difference between two sets.
Accuracy is defined as**:**

$$accuracy(h) = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{|\,h(x_i)\cap L_i\,|}{|\,h(x_i)\cup L_i\,|} \qquad (13)$$

Precision is defined as**:**

$$precision(h) = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{|\,h(x_i)\cap L_i\,|}{|\,h(x_i)\,|} \qquad (14)$$

Recallis defined as:

$$recall(h) = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{|\,h(x_i)\cap L_i\,|}{|\,L_i\,|} \qquad (15)$$

F-measure is the mean between precision and recall, which deals with a problem ofimbalanced label representation and is defined as:

$$F = \frac{1}{m_t}\sum_{i=1}^{m_t}\frac{2\times|h(x_i)\cap L_i|}{|h(x_i)|+|L_i|} \quad (16)$$

Subset Accuracy is defined as:

$$subset\_accuracy(h) = \frac{1}{m_t}\sum_{i=1}^{m_t} I(h(x_i) = L_i) \quad (17)$$

Table 1
Compare Result of ML-KNN and Threshold ML-KNN

Classifier	Accuracy	Precision	Recall	F-measure	Hamming Loss	Subset Accuracy
ML-KNN	21.30	23.08	21.35	21.90	0.38	80.49
Threshold ML-KNN	44.36	47.48	45.23	45.68	0.33	59.55

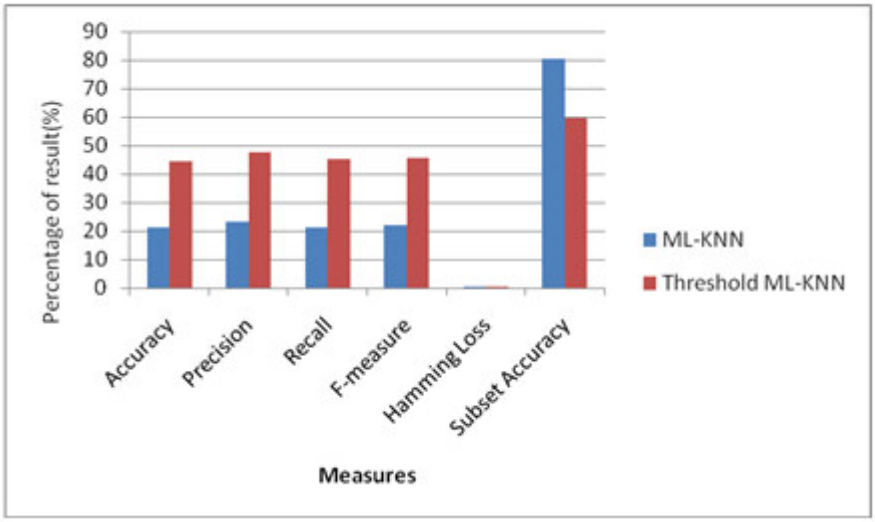

Figure 3
Result of Evaluation Criteria

By comparing the experimental results, weconclude that the threshold multi-label classifier achieves the best effect for feature training set. From Table1 It can be seen that Threshold MLKNN has given best results for plant protein dataset. After the training process, the estimated F-measure values can be used to choose the best Threshold MLKNN for each class.

Prediction Accuracy
The prediction quality was calculated by the overall prediction accuracy and prediction accuracy for each location.

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{k} p(d)}{N} \quad (18)$$

$$\text{Accuracy} = \frac{p(d)}{obs(d)} \quad (19)$$

Where N is the total number of sequences, k is the class number, obs(d) is the number of sequences observed in location d and p(d) is the number of correctly predicted sequences in location d. During testing process, only the sequences of proteins in each location were used as inputs in order to make the comparison between the two prediction classifiers under exactly the same condition. In testing process, using the equations (18 & 19) we got prediction accuracy of the ML-KNN is 63.7% and the Threshold-MLKNN prediction accuracy is 78.1%.

Table2
Prediction Accuracy Comparison

Subcellular location	Prediction accuracy	
	MLKNN	Threshold ML-KNN
Cell membrane	42.9	57.9
Cell wall	25.0	32.5
Chloroplast	86.7	88.6
Cytoplasm	39.6	47.1
Endoplasmic reticulum	40.5	49.3
Extracellular	13.6	25.7
Golgi apparatus	28.6	38.1
Mitochondrion	76.0	80.6
Nucleus	89.5	91.0
Peroxisome	66.7	73.1
Plastid	10.3	23.5
Vacuole	50.0	55.2

Results listed in Table2 are obtained with ML-KNN and Threshold MLKNN on the plant protein dataset D and we can see the overall prediction accuracy success rate achieved by Threshold MLKNN is over 78.1%, which is about 15% higher than ML-KNN.

CONCLUSION

Prediction of multi sites protein subcellular localization is a challenging problem, particularly when the system concerned contains both single and multiple sites proteins. The MLKNN algorithm is a typical multi-label k nearest neighbouring algorithm and it used for classification on several datasets but due to noise in the datasets it performs poorly. Threshold MLKNN solves this problem by taking a threshold value and it uses several methods such as Hamming Loss, accuracy, precision, recall, F-measure and subset accuracy to evaluate the performance of prediction accuracy.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Chou KC. Prediction of protein structural classes and subcellular locations. Current protein and peptide science. 2000 Sep 1;1(2):171-208.
2. Murphy RF, Boland MV, Velliste M. Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images. InISMB 2000 Aug 19; 2(8): 251-259.
3. Glory E, Murphy RF. Automated subcellular location determination and high-throughput microscopy. Developmental cell. 2007 Jan 31;12(1):7-16.
4. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular Biology of the Cell. 3rd ed. Garland. New York. 1994.p.864-66.
5. Shen HB, Chou KC. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochemical and biophysical research communications. 2005 Nov 25;337(3):752-6.
6. Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T. A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Science. 2005 Nov 1;14(11):2804-13.
7. Zhou GP, Cai YD. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. PROTEINS: Structure, Function, and Bioinformatics. 2006 May 15;63(3):681-4.
8. Emanuelsson O, Nielsen H, Brunak S, Von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J.Mol.Biol. 2000 Jul 21;300(4):1005-16.
9. Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics. 2004 Jun 1;4(6):1581-90.
10. Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Analytical biochemistry. 2007 Nov 1;370(1):1-6.
11. Vipsita S, Shee BK, Rath SK. An efficient technique for protein classification using feature extraction by artificial neural networks. InIndia Conference (INDICON), 2010 Annual IEEE 2010 Dec 17;p. 1-5. IEEE.
12. Huang DS, Zhao XM, Huang GB, Cheung YM. Classifying protein sequences using hydropathy blocks. Pattern recognition. 2006 Dec 31;39(12):2293-300.
13. Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J.* Bio*.* Chem. 2004 May 28;279(22):23262-6.
14. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001 May 15;43(3):246-55.
15. Fan GL, Li QZ. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J Theor Biol. 2012 Jul 7;304:88-95.
16. Huang C, Yuan JQ. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's

pseudo amino acid compositions. J Theor Biol. 2013 Oct 21;335:205-12.
17. Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition. 2007 Jul 31;40(7):2038-48.
18. Zhang S, Zhang HX. Modified KNN algorithm for multi-label learning. Application Research of Computers. 2011;28(12):4445-6.
19. Duan Z, Cheng JX, Zhang L. Research on multi-label learning method based on covering. Computer Engineering and Applications. 2010;46(14):20-3.
20. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research. 2000 Jan 1;28(1):45-8.
21. Łukasik M, Kuśmierczyk T, Bolikowski Ł, Nguyen HS. Hierarchical, multi-label classification of scholarly publications: modifications of ML-KNN algorithm. InIntelligent Tools for Building a Scientific Information Platform 2013;343-363. Springer Berlin Heidelberg.

Reviewers of this article

Dr.K.R. Manjula

SAP, SCE, School of Computing,
SASTRA University,
Thiumalaisamudram, Thanjuvur,
Tamilnadu - 613401, India

Prof. Srawan Kumar G.Y

Associate Professor,Nalanda Institute of Pharmaceutical Sciences, Sattenapalli, Guntur, Andrapradesh, India

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript