Ensemble classifier for protein fold pattern recognition

Hong-Bin Shen\(^1,\ast\) and Kuo-Chen Chou\(^1,2\)

\(^1\)Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China and \(^2\)Gordon Life Science Institute, San Diego, CA 92130, USA

ABSTRACT

Motivation: Prediction of protein folding patterns is one level deeper than that of protein structural classes, and hence is much more complicated and difficult. To deal with such a challenging problem, the ensemble classifier was introduced. It was formed by a set of basic classifiers, with each trained in different parameter systems, such as predicted secondary structure, hydrophobicity, van der Waals volume, polarity, polarizability, as well as different dimensions of pseudo-amino acid composition, which were extracted from a training dataset. The operation engine for the constituent individual classifiers was OET-KNN (optimized evidence-theoretic \(k\)-nearest neighbors) rule. Their outcomes were combined through a weighted voting to give a final determination for classifying a query protein. The recognition was to find the true fold among the 27 possible patterns.

Results: The overall success rate thus obtained was 62% for a testing dataset where most of the proteins have <25% sequence identity with the proteins used in training the classifier. Such a rate is 6–21% higher than the corresponding rates obtained by various existing NN (neural networks) and SVM (support vector machines) approaches, implying that the ensemble classifier is very promising and might become a useful vehicle in protein science, as well as proteomics and bioinformatics.

Availability: The ensemble classifier, called PFP-Pred, is available as a web-server at http://202.120.37.186/bioinf/fold/PFP-Pred.htm for public usage.

Contact: lifesci-sjtu@san.rr.com

Supplementary information: Supplementary data are available on Bioinformatics online.

INTRODUCTION

The avalanche of protein sequences generated in the post-genomic era has challenged us for developing computational methods by which the structural information can be timely extracted from sequence databases. Although the direct prediction of the three-dimensional (3D) structure of a protein from its sequence based on the least free energy principle is scientifically quite sound and some encouraging results already obtained in elucidating the handedness problems and packing arrangements in proteins (see e.g. Chou and Carlacci, 1991; Chou et al., 1982, 1984, 1990), it is very difficult to predict its overall fold owing to the notorious local minimum problem. Also, although it is quite successful to predict the 3D structure of a protein according to the homology modeling approach (Chou, 2004; Holm and Sander, 1999), a hurdle exists when the query protein does not have any structure-known homologous protein in the existing databases. Facing this kind of situation, can we find a different approach to predict the fold of a protein? In this paper, we shall resort to the taxonomic approach, whose underpinning is based on the assumption that the number of protein folds is limited (Chou and Zhang, 1995; Dubchak et al., 1999; Finkelstein and Ptitsyn, 1987; Murzin et al., 1995). Accordingly, predicting the 3D structure of a protein may be first converted to a problem of classification, i.e. identifying which fold pattern it belongs to. The present study was initiated in an attempt to introduce a novel approach, the ensemble classifier, to recognize the fold pattern for a query protein.

MATERIALS AND METHODS

The working (training and testing) datasets studied here were taken from Ding and Dubchak (2001). The original training dataset and testing dataset contain 313 proteins and 385 proteins, respectively. Of these proteins, however, two (i.e. 2SCMC and 2GPS) in the training dataset and two (2YHX\(_1\) and 2YHX\(_2\)) in the testing dataset do not have sequence records. These four proteins were excluded for further consideration due to lacking sequence information. Accordingly, we have 311 proteins for training dataset and 383 proteins for testing dataset. The names of the training and testing proteins and their sequences are given in Online Supplementary Materials A1 and AII, respectively. None of proteins in the testing dataset has >35% sequence identity to those in the training dataset (Ding and Dubchak, 2001). According to the SCOP database (Andreeva et al., 2004; Murzin et al., 1995), the proteins in the training and testing datasets (Online Supplementary Materials A) were further classified into the following 27-fold types (Ding and Dubchak, 2001; Dubchak et al., 1995, 1999): (1) globin-like, (2) cytochrome c, (3) DNA-binding 3-helical bundle, (4) 4-helical up-and-down bundle, (5) 4-helical cytokines, (6) EF-hand, (7) immunoglobulin-like, (8) cupredoxins, (9) viral coat and capsid proteins, (10) conA-like lectin/glucanases, (11) SH3-like barrel, (12) OB-fold, (13) beta-trefoil, (14) trypsin-like serine proteases, (15) lipocalins, (16) (TIM)-barrel, (17) FAD (also NAD)-binding motif, (18) flavodoxin-like, (19) NAD(P)h-binding Rossmann-fold, (20) P-loop, (21) thioredoxin-like, (22) ribonuclease H-like motif, (23) hydroxylases, (24) periplasmic binding protein-like, (25) beta-grasp, (26) ferredoxin-like and (27) small inhibitors, toxins, lectins. Of the above 27-fold types, types 1–6 belong to all \(a\) structural class, types 7–15 to all \(b\) class, types 16–24 to \(a+b\) class and type 25–27 to \(a+\beta\) class. Therefore, the classification of 27 folds is one level deeper than that of 4 structural classes (Cai, 2001; Chou and Zhang, 1995; Zhou, 1998; Zhou and Assa-Munt, 2001). Naturally, it is more challenging and difficult to conduct prediction among the 27-fold types than among the 4 structural classes (Chou, 1995; Chou and Maggiora, 1998).

To deal with the problem, Ding and Dubchak (2001) extracted the following six features from protein sequences: (1) amino acid composition, (2) predicted secondary structure, (3) hydrophobicity, (4) normalized van der Waals volume, (5) polarity and (6) polarizability. Of the above six features, only the amino acid composition contains 20 components, with each representing the occurrence frequency of one of the 20 native amino acids. It is much easier to handle than the other five features and has been widely used in other studies. However, their contributions are dependent on the specific 3D structures of the training proteins and hence are not generalizable. In the present study, we aimed to introduce an approach that is more generalizable and is capable of carrying a larger range of applications.
acids in a given protein (Chou and Zhang, 1994; Zhou and Doctor, 2003). For
the other five features, each contains 3 + 3 + 5 × 3 = 21 components, as
detailed in Ding and Dubchak (2001) and Dubchak et al. (1999). Based
on these multiple parameter sets and majority voting rule trained by the
proteins in the training dataset, an overall success rate of 56% was reported
(Ding and Dubchak, 2001) in predicting the fold type for the proteins in
the testing dataset.

In the present study, in order to avoid completely ignoring the
sequence-order effects, the pseudo-amino acid composition (Chou, 2001)
was used to replace the conventional amino acid composition (Chou and
Zhang, 1993; Nakashima et al., 1986) as used in (Ding and Dubchak, 2001).
However, rather than using a combined correlation function (Chou, 2001),
here the alternate correlation function between hydrophobicity and
hydrophilicity (Chou, 2005; Chou and Cai, 2005) is adopted to reflect the
sequence-order effects. For reader’s convenience, a brief introduction
about amphiphilic pseudo-amino acid composition (PseAA) is given
below.

Suppose a protein P with a sequence of L amino acid residues:
\[ R_1; R_2; R_3; R_4; R_5; R_6; \ldots; R_L, \]
where \( R_1 \) represents the residue at chain position 1, \( R_2 \) at position 2 and
so forth. The hydrophobicity and hydrophilicity of the constituent amino
acids in a protein play a very important role to its folding; e.g. many helices
in proteins are amphiphilic that is formed by the hydrophobic and
hydrophilic amino acids according to a special order along the helix
chain, as illustrated by the ‘wenxiang’ diagram (Chou et al., 1997).
Therefore, these two indices may be one of the optimal choices to reflect the
sequence-order effects. In view of this, the sequence-order effects can be
indirectly and partially, but quite effectively, reflected through the following
equations (Fig. 1):

\[
\begin{align*}
\tau_1 &= \frac{1}{L} \sum_{i=1}^{L-1} H_{i,i+1}^1, \\
\tau_2 &= \frac{1}{L} \sum_{i=1}^{L-1} H_{i,i+1}^2, \\
\tau_3 &= \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^1, \\
\tau_4 &= \frac{1}{L-2} \sum_{i=1}^{L-2} H_{i,i+2}^2, \\
\tau_{2k-2} &= \frac{1}{L-k} \sum_{i=1}^{L-k} H_{i,i+k}^1, \\
\tau_{2k} &= \frac{1}{L-k} \sum_{i=1}^{L-k} H_{i,i+k}^2,
\end{align*}
\]

where \( H_{i,j}^1 \) and \( H_{i,j}^2 \) are the hydrophobicity and hydrophilicity correlation
functions given by

\[
\begin{align*}
H_{i,j}^1 &= h^1(R_i) \cdot h^1(R_j), \\
H_{i,j}^2 &= h^2(R_i) \cdot h^2(R_j),
\end{align*}
\]

where \( h^1(R_i) \) and \( h^2(R_i) \) are, respectively, the hydrophobicity and
hydrophilicity values for the \( i \)th \((i = 1, 2, \ldots , L)\) amino acid in Equation (1), and
the dot (·) means the multiplication sign. In Equation (2) \( \tau_1 \) and \( \tau_2 \) are called
the 1st-tier correlation factors that reflect the sequence-order correlation
between all the most contiguous residues along a protein chain through
hydrophobicity and hydrophilicity, respectively [Figure 1(a1), (a2)],
\( \tau_3 \) and \( \tau_4 \) are the corresponding 2nd-tier correlation factors that reflect
the sequence-order correlation between all the 2nd most contiguous residues
[Figure 1(b1), (b2)], and so forth. Note that before substituting the values of
hydrophobicity and hydrophilicity into Equation (3), they were all subjected
to a standard conversion as described by the following equation:
\[
\begin{align*}
h_1(R_i) &= \frac{h^1(R_i) - \bar{h}_1^1}{\text{SD}(h_1^1)}, \\
h_2(R_i) &= \frac{h^2(R_i) - \bar{h}_2^2}{\text{SD}(h_2^2)},
\end{align*}
\]

where the symbols \( h_1^1(R_i) \) and \( h_2^2(R_i) \) represent the original hydrophobicity
value (Tanford, 1962) and hydrophilicity value (Hopp and Woods, 1981) for
amino acid \( R_i \), respectively (Table 1); \( \bar{h}_1^1 \) and \( \bar{h}_2^2 \) their
mean values over 20 native amino acids; SD\( (h_1^1) \) and SD\( (h_2^2) \) their
standard deviations.

The converted hydrophobicity and hydrophilicity values obtained by
Equation (4) will have a zero mean value over the 20 native amino acids
and will remain unchanged if going through the same conversion procedure
again. As we can see from Equations (1–4) as well as Figure 1, a considerable
amount of sequence-order information has been incorporated into the 2A
correlation factors through the hydrophobic and hydrophilic values of the
amino acid residues along a protein chain. By fusing the 2A amphiphilic
correlation factors into the classical amino acid composition, we have the
following augmented discrete form to represent a protein sample \( P \):

\[
P = \begin{bmatrix}
p_1 \\
p_2 \\
p_{20} \\
p_{20+1} \\
p_{20+4} \\
p_{20+14} \\
p_{20+2A}
\end{bmatrix},
\]
Table 1. The amino acid parameters used for deriving the amphiphilic pseudo-amino acid components [cf. Equation (4)]

<table>
<thead>
<tr>
<th>Code</th>
<th>Hydrophobicity $^a$ $h_i^1$</th>
<th>Hydrophobicity $^b$ $h_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.62</td>
<td>-0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.29</td>
<td>-1.0</td>
</tr>
<tr>
<td>D</td>
<td>-0.90</td>
<td>3.0</td>
</tr>
<tr>
<td>E</td>
<td>-0.74</td>
<td>3.0</td>
</tr>
<tr>
<td>F</td>
<td>1.19</td>
<td>-2.5</td>
</tr>
<tr>
<td>G</td>
<td>0.48</td>
<td>0.0</td>
</tr>
<tr>
<td>H</td>
<td>-0.40</td>
<td>-0.5</td>
</tr>
<tr>
<td>I</td>
<td>1.38</td>
<td>-1.8</td>
</tr>
<tr>
<td>K</td>
<td>-1.50</td>
<td>3.0</td>
</tr>
<tr>
<td>L</td>
<td>1.06</td>
<td>-1.8</td>
</tr>
<tr>
<td>M</td>
<td>0.64</td>
<td>-1.3</td>
</tr>
<tr>
<td>N</td>
<td>-0.78</td>
<td>2.0</td>
</tr>
<tr>
<td>P</td>
<td>0.12</td>
<td>0.0</td>
</tr>
<tr>
<td>Q</td>
<td>-0.85</td>
<td>0.2</td>
</tr>
<tr>
<td>R</td>
<td>-2.53</td>
<td>3.0</td>
</tr>
<tr>
<td>S</td>
<td>-0.18</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>-0.05</td>
<td>-0.4</td>
</tr>
<tr>
<td>V</td>
<td>1.08</td>
<td>-1.5</td>
</tr>
<tr>
<td>W</td>
<td>0.81</td>
<td>-3.4</td>
</tr>
<tr>
<td>Y</td>
<td>0.26</td>
<td>-2.3</td>
</tr>
</tbody>
</table>

$^a$The hydrophobicity values were taken from Tanford (1962).

$^b$The hydrophobicity values were taken from Hopp and Woods (1981).

where

$$p_u = \frac{\sum_{i=1}^{20} f_i \sum_{j=1}^{\lambda} \sum_{t=1}^{l} w_{ij} \gamma_{t}}{\sum_{i=1}^{20} \sum_{j=1}^{\lambda} \sum_{t=1}^{l} w_{ij} \gamma_{t}}, \quad (1 \leq u \leq 20)$$

and

$$p_{20+u} = \frac{\sum_{i=1}^{20} f_i \sum_{j=1}^{\lambda} \sum_{t=1}^{l} w_{ij} \gamma_{t}}{\sum_{i=1}^{20} \sum_{j=1}^{\lambda} \sum_{t=1}^{l} w_{ij} \gamma_{t}}, \quad (20 + 1 \leq u \leq 20 + 2\lambda)$$

where $f_i$ ($i = 1, 2, \ldots, 20$) are the normalized occurrence frequencies of the 20 native amino acids in the protein $P$, $\gamma_{t}$ the sequence-correlation factor computed according to Equation (2) and $w$ the weight factor. In the current study, we chose $w = 0.5$ to make the results of Equation (6) within the range easier to be handled ($w$ can be of course assigned with other values, but this would not have a big different impact to the final results). Therefore, the first 20 numbers in Equation (5) represent the classic amino acid composition, and the next $2\lambda$ discrete numbers reflect the amphiphilic sequence correlation along a protein chain. Such a protein representation is called ‘amphiphilic pseudo-amino acid composition’, which has the same form as the conventional amino acid composition, but contains much more information. It is through the $2\lambda$ pseudo-amino acid components that the sequence order of a protein chain and the distribution of the hydrophobic and hydrophilic amino acids along the chain are indirectly and partially reflected. It should be pointed out that, according to the definition of the classical amino acid composition, all its components must be $\geq 0$; it is not always true, however, for the pseudo-amino acid composition: the components corresponding to the sequence correlation factors may also be $< 0$.

In this study, the OET-KNN (optimized evidence-theoretic $k$-nearest neighbors) algorithm is adopted as the operation engine of a classifier (Shen and Chou, 2005). For reader’s convenience, a brief introduction about OET-KNN classifier and its key equations are given in Appendix A. However, quite different from the case of (Shen and Chou, 2005), now we have many different input types, such as the $(20 \times 5)\times (20+2\lambda)$D PseAA, 21D predicted secondary structure, 21D hydrophobicity, 21D normalized van der Waals volume, 21D polarity and 21D polarizability (Ding and Dubchak, 2001). Since a basic classifier is defined by one operation engine and one input type, one way to use the information from the multiple input types is to combine the above 6 input types into one and use a $[(21 \times 5)\times (20+2\lambda)]D$ vector to represent it. However, doing so would introduce too many parameters into the input, thereby reducing the cluster-tolerant capacity (Chou, 1999) and cross-validation success rate. Furthermore, the PseAA with a different value of $\lambda$ will become a different input type. In the present study, $\lambda$ was assigned with 1, 4, 14 and 30. Therefore, we are actually facing $5 + 4 + 9 = 18$ different input types (Table 2), and have 9 basic classifiers. To deal with this situation, we shall introduce an ensemble classifier, by which not only the other five features described in (Ding and Dubchak, 2001) but also the pseudo-amino acid compositions with a set of different $\lambda$ values can be automatically fused into one prediction system.

The framework of ensemble classifier system was established by combining numerous basic classifiers together in order to reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. Illustrated in Figure 2 is the basic framework for an ensemble classifier that consists of

Fig. 2. Flowchart to show how the ensemble classifier $C$ [Equation (7)] is formed by fusing $\Omega = 9$ basic individual classifiers: $C_1, C_2, \ldots, C_9$. A colour version of this figure appears in the Supplementary data.
\( \Omega = 9 \) basic classifiers. The final output of the ensemble is the weighted fusion of the outputs produced by the nine individual classifiers, as formulated below.

Suppose the ensemble classifier \( C \) is expressed by

\[
C = C_1 \oplus C_2 \oplus C_3 \oplus \cdots \oplus C_9
\]

(7)

where \( C_1, C_2, \ldots, C_9 \) represent the nine basic OET-KNN classifiers (Appendix A) each operating on the input derived from one of the nine features listed in Table 2. i.e. classifier \( C_1 \) operates on the 22D PseAA, \( C_2 \) on the 28D PseAA, \( C_3 \) on the 48D PseAA, \( C_4 \) on the 80D PseAA, \( C_5 \) on the 21D predicted secondary structure, \( C_6 \) on the 21D hydrophobicity, \( C_7 \) on the 21D normalized van der Waals volume, \( C_8 \) on the 21D polarity, and \( C_9 \) on the 21D polarizability. In Equation (7) the symbol \( \oplus \) denotes the fusing operator. For reader’s convenience, the values of the nine input parameter systems (cf. Table 2) for each of the proteins in the training and testing datasets are given in the Online Supplementary Materials BI and BII, respectively.

Thus, the process of how the ensemble classifier \( C \) works by fusing the nine basic classifiers \( C(i) (i=1,2,\ldots,9) \) can be formulated as follows. Suppose

\[
Y_j = \sum_{i=1}^{9} w_i \mathbb{R}_i(P,S_i) (j=1,2,\cdots,27)
\]

(8)

where \( S_1 \) is the set only containing proteins of fold type 1, \( S_2 \) the set of fold type 2, and so forth; \( \mathbb{R}_i(P,S_i) \) is the belief function or supporting degree for \( P \) belonging to \( S_i \) obtained by the \( i \)th basic classifier as defined by Equation (A5) in Appendix A; and \( w_i \) is the weighted factor, which was assigned in this study with the value of the success rate obtained by the \( i \)th basic classifier, \( C_i \), as will be further discussed below.

Thus the query protein \( P \) is predicted belonging to the fold type with which its score of Equation 8 is the highest; i.e. suppose

\[
Y_\mu = \text{Max}\{Y_1, Y_2, \cdots, Y_{27}\}
\]

(9)

where the operator \( \text{Max} \) means taking the maximum one among those in the brackets, and the subscript \( \mu \) is the very fold type predicted for the query protein \( P \). If there is a tie, the query protein may not be uniquely determined and will be randomly assigned among those with a tie, but cases like that rarely occur.

RESULTS AND DISCUSSION

To demonstrate the power of the ensemble classifier, predictions were conducted based on the same training and testing datasets used by the previous investigators (Chung and Huang, 2003; Ding and Dubchak, 2001). None of proteins in these datasets has \( >35\% \) sequence identity to any other, and most of proteins in the testing dataset have \( <25\% \) sequence identity with those in the training dataset (Ding and Dubchak, 2001). The overall success rate in recognizing the fold among the 27 folding types by the ensemble classifier for the 383 proteins in the independent dataset is given in Table 3, where, for facilitating comparison, the success rates by the individual outcomes are combined in some way, typically through a weighted voting, to give a final determination in classifying a query sample. The current ensemble classifier consists of nine basic individual classifiers. Their operation engine was OET-KNN algorithm, but they were each trained in nine different parameter systems extracted from the training dataset; i.e. 22D PseAA, 28D PseAA, 48D PseAA, 80D PseAA, 21D predicted secondary structure, 21D hydrophobicity, 21D normalized van der Waals volume, 21D polarity and 21D polarizability.

It is instructive to note that although the operation engine adopted here for the basic classifiers is the OET-KNN algorithm, others, such as the covariant discriminant algorithm and SVM algorithm, can also be used to replace the OET-KNN for forming different ensemble classifiers. Moreover, the constituent individual basic classifiers can be driven by completely different operation engines as well, and an ensemble classifier thus formed would become one with a mixture of operation engines. Similarly, we can also design an ensemble classifier by fusing both different input types and different operation engines. It is shown thru the present study that the ensemble classifier formed by fusing different input types, particularly different dimensions of pseudo-amino acid composition [(cf. Equation (5)]), is very promising for enhancing the success rate in recognizing the fold type of proteins.

APPENDIX A

The optimized evidence-theoretic \( k \)-nearest neighbors (OET-KNN) classifier

For reader’s convenience, a brief introduction of the OET-KNN classifier is given below. For further explanation, refer to (Shen and Chou, 2005). Let us consider a problem of classifying \( N \) entities into 27 classes (fold types), which can be formulated as

Table 3. Overall success rates by different approaches in recognizing the fold types for proteins in the independent testing dataset

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Success rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP (Multi-Layer Perceptron)</td>
<td>48.8</td>
</tr>
<tr>
<td>GRNN (General Regression Neural Networks)</td>
<td>44.2</td>
</tr>
<tr>
<td>SVM (Support Vector Machines)</td>
<td>45.2</td>
</tr>
<tr>
<td>SVM* (Ding and Dubchak, 2001)</td>
<td>51.1</td>
</tr>
<tr>
<td>SVM** (Ding and Dubchak, 2001)</td>
<td>56.0</td>
</tr>
<tr>
<td>Ensemble Classifier</td>
<td>62.1</td>
</tr>
</tbody>
</table>

*The training method for NN is ‘one against others’.
*The training method for SVM is ‘one against others’.
*The training method for SVM is ‘unique one against others’.
*The training method for SVM is ‘all against all’.
*The ensemble classifier is constructed by nine OET-KNN classifiers [cf. Equation (7)], and the number of neighbors in each OET-KNN classifier is 8.

CONCLUSIONS

An ensemble classifier is formed by a set of basic classifiers, whose individual outcomes are combined in some way, typically through a weighted voting, to give a final determination in classifying a query sample. The current ensemble classifier consists of nine basic individual classifiers. Their operation engine was OET-KNN algorithm, but they were each trained in nine different parameter systems extracted from the training dataset; i.e. 22D PseAA, 28D PseAA, 48D PseAA, 80D PseAA, 21D predicted secondary structure, 21D hydrophobicity, 21D normalized van der Waals volume, 21D polarity and 21D polarizability.

It is instructive to note that although the operation engine adopted here for the basic classifiers is the OET-KNN algorithm, others, such as the covariant discriminant algorithm and SVM algorithm, can also be used to replace the OET-KNN for forming different ensemble classifiers. Moreover, the constituent individual basic classifiers can be driven by completely different operation engines as well, and an ensemble classifier thus formed would become one with a mixture of operation engines. Similarly, we can also design an ensemble classifier by fusing both different input types and different operation engines. It is shown thru the present study that the ensemble classifier formed by fusing different input types, particularly different dimensions of pseudo-amino acid composition [(cf. Equation (5)]), is very promising for enhancing the success rate in recognizing the fold type of proteins.
The available information is assumed to consist of a training dataset
\[ N = \{(P_1, \theta_1), \ldots, (P_N, \theta_N)\} \] (A2)
where the \( N \) entities \( P_i (i = 1, \ldots, N) \) and their corresponding pattern (class) labels \( \theta_i (i = 1, \ldots, N) \) take values in \( \mathcal{F} \) of Equation (A1). According to the KNN (k-nearest neighbors) rule (Cover and Hart, 1967), an unclassified entity \( P \) is assigned to the class represented by a majority of its \( K \)-nearest neighbors of \( P \). Owing to its good performance and simple-to-use feature, the KNN rule, also named as ‘voting KNN rule’, is quite popular in pattern recognition community.

The ET-KNN (evidence theoretic \( k \)-nearest neighbors) rule is a pattern classification method based on the Dempster–Shafer theory of belief functions (Denœux, 1995). In the classification process, each neighbor of a pattern to be classified is considered as an item of evidence supporting certain hypotheses concerning the class membership of that pattern. Based on this evidence, basic belief masses are assigned to each subset concerned. Such masses are obtained for each of the \( k \)-nearest neighbors of the pattern under consideration and aggregated using the Dempster’s rule of combination (Shafer, 1976). A decision is made by assigning a pattern to the class with the maximum credibility.

Suppose \( P \) is a query protein to be classified, and \( S_P \) is the set of its \( k \)-nearest neighbors in the training dataset \( N \) of Equation (A2). Thus, for any \( P_i \in S_P \), the knowledge that \( P_i \) belongs to class \( \Phi_\mu \in \mathcal{F} \) can be considered as a piece of evidence that increases our belief that \( P \) also belongs to \( \Phi_\mu \). According to the basic belief assignment mapping theory (Shafer, 1976), this item of evidence can be formulated by
\[
\mathbb{R}(P, \Phi_\mu) = \alpha_0 \exp \left[ -\gamma_\mu D^2(P, P_i) \right] 
\] (A3)
where \( \alpha_0 \) is a fixed parameter, \( \gamma_\mu \) is a parameter associated with class \( \Phi_\mu \) and \( D^2(P, P_i) \) is the square Euclidean distance between \( P \) and \( P_i \). In the ET-KNN rule, it was not addressed how to optimally select the parameters. In 1998 an optimization procedure to determine the optimal or near-optimal parameter values was proposed from the data by minimizing an error function (Zouhal and Denœux, 1998). It was observed that the OET-KNN rule obtained thru such an optimization treatment would lead to a substantial improvement in classification accuracy. The optimal parameter thus obtained for \( \alpha_0 \) of Equation A3 is 0.95, and those for \( \gamma_\mu \) are given in Table A1.

Table A1. The optimal parameter \( \gamma_j \) \((j = 1, \ldots, 27)\) in Equation A3 obtained thru the optimized procedure (Zouhal & Denœux, 1998) for the 9 basic individual classifiers in Equation 7

<table>
<thead>
<tr>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( \gamma_4 )</th>
<th>( \gamma_5 )</th>
<th>( \gamma_6 )</th>
<th>( \gamma_7 )</th>
<th>( \gamma_8 )</th>
<th>( \gamma_9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1028</td>
<td>0.2908</td>
<td>0.0656</td>
<td>0.0888</td>
<td>0.0798</td>
<td>0.0931</td>
<td>0.0954</td>
<td>0.1241</td>
<td>0.1476</td>
</tr>
<tr>
<td>0.0714</td>
<td>0.0727</td>
<td>0.0490</td>
<td>0.0480</td>
<td>0.0525</td>
<td>0.1176</td>
<td>0.0783</td>
<td>0.1013</td>
<td>0.1076</td>
</tr>
<tr>
<td>0.0398</td>
<td>0.0523</td>
<td>0.0240</td>
<td>0.0752</td>
<td>0.0274</td>
<td>0.0468</td>
<td>0.0520</td>
<td>0.0475</td>
<td>0.0700</td>
</tr>
<tr>
<td>0.0434</td>
<td>0.1324</td>
<td>0.0435</td>
<td>0.0597</td>
<td>0.0312</td>
<td>0.0581</td>
<td>0.0543</td>
<td>0.0462</td>
<td>0.0699</td>
</tr>
<tr>
<td>0.4848</td>
<td>0.0585</td>
<td>0.0469</td>
<td>0.4306</td>
<td>0.4146</td>
<td>0.1298</td>
<td>0.2871</td>
<td>0.1890</td>
<td>0.1386</td>
</tr>
<tr>
<td>0.1275</td>
<td>0.0784</td>
<td>0.0604</td>
<td>0.1541</td>
<td>0.1510</td>
<td>0.4586</td>
<td>0.0973</td>
<td>0.3587</td>
<td>0.1366</td>
</tr>
<tr>
<td>0.1977</td>
<td>0.3654</td>
<td>0.2482</td>
<td>0.4092</td>
<td>0.0866</td>
<td>0.0981</td>
<td>0.3723</td>
<td>0.5057</td>
<td>0.0910</td>
</tr>
<tr>
<td>0.1396</td>
<td>0.2218</td>
<td>0.2988</td>
<td>0.4026</td>
<td>0.0543</td>
<td>0.1741</td>
<td>0.1200</td>
<td>0.1365</td>
<td>0.0425</td>
</tr>
<tr>
<td>0.0682</td>
<td>0.0387</td>
<td>0.0445</td>
<td>0.1741</td>
<td>0.1407</td>
<td>0.0981</td>
<td>0.0385</td>
<td>0.0395</td>
<td>0.0425</td>
</tr>
</tbody>
</table>

1721
The belief function of $P$ belonging to class $\Phi_\mu$ is a combination of its $k$-nearest neighbors, and can be formulated as

$$\text{Bel}(P, \Phi_\mu) = \left( \prod_{i=1}^k \text{Bel}(P_i, \Phi_\mu) \right) \text{Bel}(P, \Phi_\mu)$$

where $\Phi_\mu$ is called the orthogonal sum, which is commutative and associative. According to Dempster’s rule (Shafer, 1976), the belief function of Equation A4 can be expressed as

$$\text{Bel}(P, \Phi_\mu) = \frac{\sum_{x \subseteq \Phi_\mu} \text{Bel}(P, \Phi_x) \text{Bel}(P, \Phi_\mu \setminus x)}{1 - \sum_{x \subseteq \Phi_\mu} \text{Bel}(P, \Phi_x) \text{Bel}(P, \Phi_\mu \setminus x)}$$

where $\Phi_x$ is the $i$-th possible subset of $\Phi_\mu$, and $\subseteq$, $\cap$, and $\emptyset$ are the symbols in set theory, representing 'contained in', 'intersection', and the empty set, respectively.

A decision is made by assigning the query protein $P$ to the class with which the belief or credibility function of Equation A5 has the maximum value; i.e. if

$$\text{Bel}(P, \Phi_\mu) = \max \left\{ \text{Bel}(P, \Phi_1), \text{Bel}(P, \Phi_2), \ldots, \text{Bel}(P, \Phi_m) \right\}$$

where $\mu = 1, 2, \ldots, 27$ and the operator $\max$ means taking the maximum one among those in the brackets, then the class $\Phi_\mu$ is the class predicted for the query protein.

Conflict of Interest: none declared.

REFERENCES


