Expression Detection Based on a Novel Emotion Recognition Method

Xun Gong *
School of Information Science and Technology, Southwest Jiaotong University,
Chengdu 610031, P. R. China

Yong Yang
Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications,
Chongqing 400065, P. R. China
E-mail: yangyong@cqupt.edu.cn
http://www.cqupt.edu.cn/

Jianhui Lin
Traction Power State Key Laboratory, Southwest Jiaotong University,
Chengdu 610031, P. R. China
E-mail: linjhyz@126.com

Tianrui Li
School of Information Science and Technology, Southwest Jiaotong University,
Chengdu 610031, P. R. China
E-mail: trli30@gmail.com
http://sist.swjtu.edu.cn/iip/ui/index.aspx

Accepted: 27-02-2009
Received: 21-09-2010

Abstract

As facial expression is an essential way to convey human's feelings, in this paper, a dynamic selection ensemble learning method is proposed to analyze their emotion automatically. A feature selection algorithm is proposed at first based on rough set and the domain oriented data driven data mining theory, which can get multiple reducts and candidate classifiers. Then the nearest neighborhood of each unseen sample is found in a validation subset and the most accurate classifier is extracted from the candidate classifiers. Finally, the selected classifier is used to recognize unseen samples. Experimental results show that the proposed method is effective and suitable for emotion recognition.

Keywords: emotion recognition; ensemble learning; rough set; data mining

1. Introduction

Even there have been some commercial products, such as the amazing robots in Japan, it still takes a long time to make a computer act as a human to recognize human emotions since there are many problems remain unsolved in psychology and cognitive fields. Currently, emotion recognition is studied with such methods as ANN, fuzzy set, SVM, HMM and rough set, where the recognition rate often ranges from 64% to 98% [1–2]. To achieve a better result based on existing classification methods, ensemble learning is proposed in 1990s' [3], which construct a set of candidate classifiers and then

*Corresponding author. Tel.: +86-28-86466426. E-mail address: gongxun@foxmail.com (X. Gong)

classify new objects by integrating the prediction of the candidate classifiers. Experiments have validated that an ensemble system is much more accurate than any separate classifier. Ditterrich proved the effectiveness of ensemble methods from the viewpoint of statistic, computation and representation in Ref. 4. Currently, ensemble methods have been applied widely over a variety of areas, e.g., pattern recognition, network security and medical diagnosis [4–7]. Ensemble strategies are commonly used to gain a different subset of the original dataset by multi-sampling the training set, like bagging [6], boosting [7] and cross-validation. These methods work well especially for unstable learning algorithms, such as decision trees and neural network. Some other methods are also studied, such as manipulating the output targets [8] and injecting randomness into classifiers [9]. In addition, the ensemble feature selection (EFS) [10] is another effective approach for ensemble, which is also a classical ensemble method. It uses different feature subset as input feature to construct a candidate classifier.

In this work, we propose a novel emotion recognition method based on EFS and rough set theory. At first, a feature selection algorithm is proposed based on rough set and domain oriented data driven data mining (3DM) theory [11–12], which can obtain multiple reducts and candidate classifiers. The nearest neighborhood of each unseen sample is then found in a validation subset and the most accurate classifier is selected from the candidates. At last, the selected classifier is used to classify the unseen samples. The proposed method is proved to be effective by extensive experiments.

The remainder of this paper is structured as follows. In Section 2, the tolerance relation model for continuous value information system is introduced. The proposed emotion recognition method based on dynamic selection ensemble learning is then discussed. Simulation experiments and discussions are presented in Section 3. Finally, conclusion is drawn in Section 4.

2. A dynamic selection ensemble learning model

Rough Set (RS) is a valid mathematical theory for dealing with imprecise, uncertain, and vague information, which was developed by Professor Z. Pawlak in 1980s [13–14]. Until now, RS has been successfully used in many fields, such as machine learning, pattern recognition, intelligent data analyzing and etc. The most advantage of RS is its great ability of attribute reduction (knowledge reduction and feature selection).

In this section, RS is used as a tool for feature reduction in the proposed emotion recognition method based on EFS. For traditional RS theory, a pretreatment of discretization is necessary as the facial features are continuous. Stand on this point, information will be unavoidably lost or changed during the pretreatment and the result would be affected afterward. To solve this problem, a feature selection method based on tolerance relation for emotion recognition is proposed in this paper, in which discretization is not needed. Based on the idea of 3DM, a method for selecting suitable threshold of tolerance relation is also proposed in section 2.2.

2.1 *Tolerance relation model for continuous value information system*

Some basic RS concepts of continuous value information system are briefly described for the convenience of the following discussion.

Def. 1 A decision information system is defined as a quadruple $S = (U, C \cup D, V, f)$, where U is a finite set of objects, C is the condition attribute set and $D = \{d\}$ is the decision attribute set. $\forall c \in C$, with every attribute $a \in C \cup D$, a set of its values V_a is associated. Each attribute a determines a function $f_a : U \rightarrow V_a$.

Def. 2 For a subset of attributes $B \subseteq A$, an indiscernibility relation is defined by $Ind(B) = \{(x, y) \in U \times U : \forall_{a \in B}(a_x = a_y)\}$, in which a_x and a_y are values of the attribute a of x and y.

The indiscernibility relation defined in this way is an equivalence relation. Obviously, $Ind(B)=\cap_{b \in B} Ind(\{b\})$. By $U/Ind(B)$ we mean the set of all equivalence classes in the relation $Ind(B)$. The classical RS theory is based on an observation that objects may be indiscernible due to limited available information, and the indiscernibility relation defined in this way is an equivalence relation indeed. The intuition behind the notion of an indiscernibility relation is that selecting a set of attribute $B \subseteq A$ effectively defines a partition of the universe into sets of objects that cannot be discerned using the attributes in B only. The equivalence classes $E_i \in U/Ind(B)$, induced by a set of attributes $B \subseteq A$, are referred to as object classes or simply classes. The classes resulted from $Ind(A)$ and $Ind(D)$ are called condition classes and decision classes, respectively.

Def. 3 A continuous value decision information system is defined as a pair $S = (U, R, V, f)$, where U is a finite set of objects and $R = C \cup D$ is a finite set of attributes, C is the condition attribute set and $D = \{d\}$ is the decision attribute set. $\forall c \in C$, c is continuous attribute value, $\forall d \in D$, d is a continuous attribute or a discrete value attribute.

A facial expression information system is a continuous value decision information system according to Def. 3. Facial attributes are continuous and might be imprecise in some extent, and the process of discretization will affect the result of emotion recognition, so it is suitable to take continuous values equal to each other in some range.

Def. 4 A binary relation $R(x,y)$ defined on an attribute set B is called a tolerance relation if it satisfies:

Symmetrical: $\forall_{x,y \in U}(R(x,y) = R(y,x))$.

Reflective: $\forall_{x \in U}(R(x,x) = R(x,x))$.

A new relation for continuous value decision information systems is defined as follows:

Def. 5 Let an information system $S = (U, R, V, f)$ be a continuous value decision information system, and a new relation $R(x,y)$ be defined as:

$$R(x,y) = \{(x,y) \mid x \in U \wedge y \in U \wedge \forall_{a \in C}(\left|a_x - a_y\right| \le e, 0 \le e \le 1)\}$$

It is easy to see that $R(x, y)$ is a tolerance relation according to Def. 4 since $R(x, y)$ is symmetrical and reflective. An equivalence relation constitutes a partition of U, but a tolerance relation constitutes a cover of U, and an equivalence relation is a particular type of a tolerance relation.

Def. 6 Let $R(x, y)$ be a tolerance relation according to Def. 3, $n_R(x_i) = \{x_j \mid x_j \in U \wedge \forall_{a \in C}(\left|a_{x_i} - a_{x_j}\right| \le e)\}$ is called the tolerance class of x_i, and $\left|n_R(x_i)\right| = \left|\{x_j \mid x_j \in n_R(x_i), 1 \le j \le U\}\right|$ is the cardinality of the tolerance class of x_i.

According to Def. 6, $\forall x \in U$, bigger tolerance class of x will lead to more uncertainties and less knowledge, and vise versa. The concepts of knowledge entropy and conditional entropy are defined:

Def. 7 Let $U = \{x_1, x_2, ..., x_{|U|}\}$, $R(x_i, x_j)$ be a tolerance relation defined on an attribute set B, knowledge entropy $E(R)$ of relation R is defined as

$$E(R) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_R(x_i)|}{|U|}.$$

Def. 8 Let R and Q be tolerance relations defined on U, $R \cup Q$ is a relation satisfying R and Q simultaneous, and it is a tolerance relation too. $\forall x_i \in U$, $n_{R \cup Q}(x_i) = n_R(x_i) \cap n_Q(x_i)$, therefore, the knowledge entropy of $R \cup Q$ can be defined as $E(R \cup Q) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|n_{R \cup Q}(x_i)\right|}{|U|}$.

Def. 9 Let R and Q be tolerance relations defined on U, the conditional entropy of R with respect to Q is defined as $E(Q \mid R) = E(R \cup Q) - E(R)$.

Let $S = (U, R, V, f)$ be a continuous value decision information system, the relation K be a tolerance relation defined on its conditional attribute set C, the relation L be an equivalence relation (a special tolerance relation) defined on its decision attribute set D. From Def. 7, Def. 8 and Def. 9, we have:

$$\begin{aligned} E(D \mid C) &= E(L \mid K) = E(K \cup L) - E(K) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{K \cup L}(x_i)|}{|U|} - (-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_K(x_i)|}{|U|}) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{K \cup L}(x_i)|}{|n_K(x_i)|} \end{aligned}$$

where, the conditional entropy $E(D|C)$ has a clear meaning, *i.e.*, it is a ratio between the knowledge of all attributes (condition attribute set plus decision attribute set) and the knowledge of the condition attribute set.

2.2 *Parameter selection for tolerance relation model*

In this subsection, a novel attribute reduction algorithm is developed based on the idea of domain-oriented data-driven data mining (3DM), which is a data mining theory proposed by Wang *et al.* [11–12]. According to the idea of 3DM, knowledge can be expressed in many different ways. There should be some relationship among different formats of the same knowledge. In order to keep the knowledge unchanged in a data mining process, the properties of the knowledge should remain unchanged during the knowledge transformation process [15]. Otherwise, information losses may occur in the transformation process. Knowledge reduction can be seen as a process of knowledge transformation, during which the knowledge properties will be remained.

With the background of our application to emotion recognition, there is not a face that looks exactly the same as others and the equivalent argument could exist in facial emotion as well. If there are two different emotion samples, there must be some different features in the samples. Hence, an emotion sample belongs to an

emotion state according to its features which are different to the other one. So, the indiscernibility of conditional attribute set with respect to the decision attribute set can be taken as an important property of knowledge when knowledge reduction is performed in emotion recognition. Based on the idea of 3DM, the indiscernibility is decided when a decision information table is given, and the ability should be unchanged in the process of attribute reduction.

Def. 10 Let $S=(U, R, V, f)$ be a continuous value decision information system, if $\forall_{x_i,x_j\in U}(d_{x_i}\neq d_{x_j}\rightarrow \exists_{a\in C}(a_{x_i}\neq a_{x_j}))$, there is an indiscernibility of the conditional attribute set with respect to the decision attribute set in the continuous value decision information system *S*.

The indiscernibility can be seen as a fundamental ability that a continuous information decision system has. According to 3DM, the indiscernibility should be unchanged in the process of knowledge acquisition. Therefore, the indiscernibility should be held if feature selection is done on a continuous value decision information system based on tolerance relation. Based on the standpoint that attribute values could be equal in some range while not be equal exactly in a continuous value decision information system, according to Def. 10, $\forall_{x_i,x_j\in U}(d_{x_i}\neq d_{x_j}\rightarrow \exists_{a\in C}(|a_{x_i}-a_{x_j}|>e))$, and according to Def. 6, $x_j\notin n_R(x_i)$, $x_i\notin n_R(x_j)$, $n_R(x_j)\neq n_R(x_j)$. Therefore, the indiscernibility of a tolerance relation can be obtained.

Def. 11 Let *R(x,y)* be a tolerance relation according to Def. 5, if $\forall_{x_i,x_j\in U}(d_{x_i}\neq d_{x_j}\rightarrow n_R(x_i)\neq n_R(x_j))$, *R(x,y)* has the certain discernibility.

If *R(x,y)* has certain discernibility , according to Def. 11, $\forall_{x_i,x_j\in U}(n_R(x_i)=n_R(x_j)\rightarrow d_{x_i}=d_{x_j})$, therefore, $\forall_{x_i,x_j\in U}(x_i,x_j\in n_R(x_i)\rightarrow d_{x_i}=d_{x_j})$.

Theorem 1 In a tolerance relation, $E(D|C)=0$ is a necessary and sufficient condition if there is an indiscernibility of the conditional attribute set with respect to the decision attribute set.

Proof. Let $S=(U, R, V, f)$ be a continuous value information system, relation *K* be a tolerance relation defined on condition attribute set *C*, relation *L* be an equivalence relation (a special tolerance relation) defined on decision attribute set *D*.

(Necessity). If there is certain discernibility for the condition attribute set with respect to the decision attribute set in tolerance relation, according to Def. 11, $\forall_{x_i,x_j\in U}(x_i,x_j\in n_K(x_i)\rightarrow d_{x_i}=d_{x_j})$, then, $n_K(x_i)\subseteq n_L(x_i)$, $n_{K\cup L}(x_i)=n_K(x_i)$, $|n_{K\cup L}(x_i)|=|n_K(x_i)|$, $E(D|C)=E(L|K)$

$$=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{K\cup L}(x_i)|}{|n_K(x_i)|}=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2 1=0.$$

(Sufficiency). $\forall x_i\in U$, we have $n_{K\cup L}(x_i)\subseteq n_K(x_i)$, $|n_{K\cup L}(x_i)|\leq|n_K(x_i)|$. Since

$$E(D|C)=E(L|K)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{K\cup L}(x_i)|}{|n_K(x_i)|}=0,$$

we can have $\forall x_i\in U$, $|n_{K\cup L}(x_i)|=|n_K(x_i)|$, that is, $n_{K\cup L}(x_i)=n_K(x_i)$. Therefore, decision values should be equal for different samples included in the same tolerance class. Accordingly, we can have $\forall_{x_i,x_j\in U}(x_i,x_j\in n_R(x_i)\rightarrow d_{x_i}=d_{x_j})$. Therefore, $\forall_{x_i,x_j\in U}(d_{x_i}\neq d_{x_j}\rightarrow\exists_{a\in C}(a_{x_i}\neq a_{x_j}))$, and there is certain discernibility for condition attribute set with respect to decision attribute set in tolerance relation. This completes the proof.

From Theorem 1, we know that the *R(x,y)* has an indiscernibility that could be taken as a measurement of $E(D|C)=0$.

For a given continuous decision information system *S*, there could be many different tolerance relations through choosing different threshold *e* under the condition $E(D|C)=0$. However, the biggest granule and the best generalization for knowledge is always needed for knowledge acquisition. According to this principle, we can have the following result.

1) If the threshold *e* in a tolerance relation is 0, then the tolerance class $n_R(x_i)$ of an instance x_i only contains x_i itself, that is, $n_{R\cup Q}(x_i)=n_R(x_i)=\{x_i\}$, and $E(D|C)=0$. It is the smallest tolerance class of the tolerance relation, and it is the smallest knowledge granule and the smallest generalization.
2) If threshold *e* in a tolerance relation is increased from 0, then $n_R(x_i)$ and $n_{R\cup Q}(x_i)$ are both increased. If $n_R(x)\subseteq n_Q(x)$, then $n_{R\cup Q}(x_i)=n_R(x_i)$, $|n_{R\cup Q}(x_i)|=|n_R(x_i)|$, $E(D|C)=0$, and knowledge granule is increased.
3) If threshold *e* in tolerance relation is increased to a critical point named e_{opt}. $n_R(x_i)$ and $n_{R\cup Q}(x_i)$ are both increased, and $n_{R\cup Q}(x_i)=n_R(x_i)$, $|n_{R\cup Q}(x_i)|=|n_R(x_i)|$, $E(D|C)=0$, and knowledge granule is the biggest under the condition discernable-ability is unchanged for conditional attribute set with respect to decision attribute set in tolerance relation.

4) If threshold e in a tolerance relation is increased from e_{opt} and $e<1$, then $n_{R\cup Q}(x_i)\neq n_R(x_i)$, $|n_{R\cup Q}(x_i)|\neq|n_R(x_i)|$, $E(D|C)\neq 0$, then the discernable-ability is changed. If $\forall_{x_i\in U}(n_Q(x_i)\subset n_R(x_i))$, then $n_{R\cup Q}(x_i)=n_Q(x_i)$, $|n_{R\cup Q}(x_i)|=|n_Q(x_i)|$, and $|n_Q(x_i)|<|n_R(x_i)|$. So, $E(D|C)=E(Q|R)=$

$$-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{R\cup Q}(x_i)|}{|n_R(x_i)|}=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_Q(x_i)|}{|n_R(x_i)|}>0,$$

Since $|n_Q(x_i)|$ is hold and $|n_R(x_i)|$ is increased with the threshold of e increase, $E(D|C)$ is increased.

5) If threshold e in a tolerance relation is increased to $e=1$, then $n_R(x_i)=U$ and $\forall_{x_i\in U}(n_Q(x_i)\subseteq n_R(x_i))$, $n_{R\cup Q}(x_i)=n_Q(x_i)$, $|n_{R\cup Q}(x_i)|=|n_Q(x_i)|$, so, $E(D|C)=E(Q|R)=$

$$-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_{R\cup Q}(x_i)|}{|n_R(x_i)|}=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|n_Q(x_i)|}{|U|}.$$

Since the equivalence class of Q is hold, $E(D|C)$ is constant.

The relationship between entropy, condition entropy and e is shown in Fig. 1.

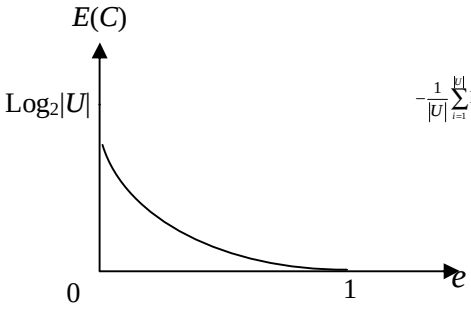


Fig 1.a Relationship between $E(C)$ and e

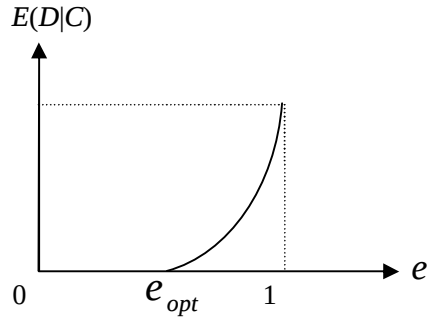


Fig 1.b Relationship between $E(D|C)$ and e

From Fig. 1 and the discussion above, if the threshold value of e is e_{opt}, it could make $E(D|C)=0$ and the classification ability of the conditional attribute set with respect to decision attribute set is unchanged. At the same time, the tolerance class of x is the biggest with $E(D|C)=0$. The knowledge granule of the conditional attribute set is the biggest in the case of e_{opt}, the knowledge generalization is the best too.

Based on the discussion above, the suitable threshold (e_{opt}) is found and the tolerance relation is set up accordingly.

2.3 *An algorithm for finding candidate classifiers of ensemble*

Based on the tolerance relation model proposed above, a new algorithm for finding the core of a decision table is proposed as follows.

Alg. 1 The algorithm for finding core

Input: a continuous value decision information system $S=(U, C\cup D, V, f)$, where U is a finite set of objects, C is the condition attribute set and $D=\{d\}$ is the decision attribute set.

Output: the core $Core_D(C)$ of S

Step1 Compute e_{opt}, then set up a tolerance relation model.

Step2 $Core_D(C)=\varnothing$.

Step3 $\forall a_i\in C$, If $E(D|C)<E(D|C-\{a_i\})$, then $Core_D(C)=Core_D(C)\cup\{a_i\}$

Step 4 Reture $Core_D(C)$

After calculating the core of S according to the Algorithm 1, an algorithm for finding multiple reducts of S is proposed as follows.

Alg. 2 An algorithm for computing multiple reducts.

Input: a continuous value decision information system $S=(U, C\cup D, V, f)$

Output: a set of reducts $\cup_i REDU_i$,

Step1 Compute the core $Core_D(C)$ of decision table S using Algorithm 1.

Step2 $AR=C-Core_D(C)$, $REDU_i=Core_D(C)$, $i=1$.

Step3 $\forall a_i\in AR$, compute $E(D|\{a_i\})$, and sort AR by $E(D|\{a_i\})$ ascendly.

Step4 While the attributes in $\cup_i REDU_i$ do not

include all the attributes in C

Step 4.1 **While** ($E(D \mid REDU_i) \neq E(D \mid C)$)

$\forall a_j \in AR,\ REDU_i = REDU_i + a_j$,

$AR = AR - a_j$

Compute $E(D \mid REDU_i)$

If ($E(D \mid REDU_i) \neq E(D \mid C) \wedge REDU_i$ - $CORE_D(C) == AR$)

i=i-1,

Goto step5

Endif

Endwhile

Step 4.2 $N = |REDU_i|$

Step 4.3 **For** j=0 to N-1

If $a_j \in REDU_i$ and $a_j \notin CORE$ then

$REDU_i = REDU_i - a_j$, get $E(D \mid REDU_i)$

If $E(D \mid REDU_i) \neq E(D \mid C)$ then

$REDU_i = REDU_i + a_j$

Endif

Endif

Endfor

Step 4.4 $AR = AR - a$,

$a \in REDU_i \wedge a = \min(E(D \mid \{a_j\}), a_j \in REDU_i$

Step 4.5 $i = i + 1$

Step 5 Reture $\cup_i REDU_i$

Algorithm 2 could find multiple reducts of a decision table. Therefore, all the candidate classifiers could be generated accordingly. In this paper, SVM is used as the classifier, and all the classifiers take the same parameters.

2.4 *Dynamic ensemble static selection*

There are different ways for ensemble. Selective ensemble is a popular one. It selects the most diversity classifiers and integrates predictions of the selective classifiers. Unfortunately, it is difficult to define the measure of the diversity in real applications. In this paper, a dynamic selection method is used instead of the statically selective method.

Alg. 3 An algorithm of dynamic selection EFS.

Input: A decision table $S = (U, C \cup D, V, f)$, and training subset, validation subset and testing subset.

Output: Classification output of the ensemble.

Step1 Find multiple reducts of the training subset using Algorithm 2, and train all the candidate classifiers.

Step2 For each sample ***x*** in the test subset, do:

For each reduct

Calculate the ***K*** nearest neighborhood in the validation subset.

Classify the ***K*** nearest neighborhood by the candidate classifiers.

Step3 Classify ***x*** using the classifier with the highest correct classification ratio in Step 2, Reture classification result as the output of the ensemble.

3. Experiments and discussion

3.1 *Testing dataset*

Three facial emotional datasets are used for tests: one comes from the Cohn-Kanade AU-Coded Facial Expression (CKACFE) database [16] and the dataset is more representative of Caucasian to some extent. Another one is the Japanese female facial expression (JAFFE) database [17] and it is more representative of Asian women. The third one named CQUPTE [18] is collected from 8 Chinese graduate students (four female and four male). Details of the datasets are listed in Table 1.

Some examples are shown in Fig. 2. They are of the emotion of happiness, sadness, fear, disgust, surprise, angry from left to right. Each dataset are split into a training subset, a validation set and a test set of the ratio of 6:1:1, and 8-fold cross-validation are taken.

Facial expression of human being is expressed by the shape and position of facial components such as eyebrows, eyes, mouth, nose, etc. The geometric features, appearance features, wavelet features and mixture features of facial are popular for emotion recognition in recent years. The geometric facial features represent the shape and locations of facial components, and it is used in the experiments since it is obvious and intuitive for the facial expression. The geometric facial features are the distance between two different feature points which are according to a defined criterion. The MPEG-4 standard is a popular standard for feature point selection. It extends facial action coding system (FACS) to derive facial definition parameters (FDP) and facial animation parameters (FAP). There are 68 FAP parameters, in which 66 low parameters are defined according to FDP parameters to describe the motion of a human face. The FDP and low

level FAP can constitute a concise representation of a face, and they are adequate for basic emotion recognition because of the varieties of expressive parameter. In the experiments, 52 low FAP parameters are chosen to represent emotion because some FAP parameters have little effect on facial expression. For example, the FAP parameter named raise_l_ear, which denote the vertical displacement of left ear. Thus, a feature point set including 52 feature points is defined as shown in Fig. 3. Based on the feature points, 33 facial features are extracted for emotion recognition according to Ref. 19 and listed in Table 2. 33 facial features can be divided into three groups. There are 17 features in the first group which concern eyes and consist of $d_0, d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_{16}, d_{17}, d_{19}, d_{20}, d_{25}, d_{26}, d_{27}, d_{28}$ and d_{29}; there are 6 features in the second group which concern cheek and consist of $d_9, d_{10}, d_{18}, d_{21}, d_{30}$ and d_{31}; there are 10 features in the third group which concern mouth and consist of $d_8, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{22}, d_{23}, d_{24}$ and d_{32}. In Table 2, A is the midpoint of point 19 and 23, and B is the midpoint of point 27 and 31. dis(i, j) denotes the Euclidian distance between point i and j; hei(i, j) denotes the horizontal distance between point i and j; wid(i, j) denotes the vertical distance between i and j. Since the distance between point 23 and 27 is stable for all kinds of expression, we normalize the distance features in the following way:

Firstly, $x_i' = d_i / d$, i=0, 1, ..., 32, d is the distance between point 23 and 27. Secondly, the normalized distance is calculated using the following formula:

$$x_i = \frac{x_i' - min(x_i')}{max(x_i') - min(x_i')}, \ x_i \in [0, 1].$$

3.2 *Algorithm evaluation*

Here, three algorithms are evaluated for comparison: (1) the proposed method DEFS; (2) all the classifiers are trained according to Algorithm 2, and the output of all the classifiers are combined according to the criterion of majority voting; (3) a reduction algorithm based on conditional entropy in rough set theory—conditional entropy based algorithm for reduction of knowledge without computing core (CEBARKNC) —proposed by Guoyin Wang is used for feature selection method in emotion recognition, and SVM is taken as classify [20]; (4) a feature selection algorithm proposed by Xiaohu Hu is used for feature selection method in emotion recognition, and SVM is taken as classify too [21]. SVM with same parameters are taken as classifiers in these comparative experiments. 8-fold cross-validation is taken for all the experiments.

Results of the comparative experiments are shown in Table 3, where 'n' is the number of classifiers. Through comparing all the three comparative experiment results, we can find that the proposed method is most accurate. Therefore, we can draw a conclusion that the proposed method is effective and a suitable method for emotion recognition.

Table 1 Three facial emotional datasets.

Dataset Name	Samples	People	Emotion classes
CKACFE	405	97	Happiness, Sadness, Surprise, Anger, Disgust, Fear, Neutral
JAFFE	213	10	Happiness, Sadness, Surprise, Anger, Disgust, Fear, Neutral
CQUPTE	652	8	Happiness, Sadness, Surprise, Anger, Disgust, Fear, Neutral

(a) Examples of CKACFE database

(b) Examples of JAFFE database

(c) Examples of CQUPTE database

Fig. 2 Facial emotion samples

Table 2 33 features defined on 52 feature points

feature	description	feature	description	feature	description
d0	dis(11,19)	d11	dis(39,44)	d22	dis(44,48)/2
d1	dis(18,31)	d12	dis(39,48)	d23	dis(45,51)
d2	dis(21,25)	d13	dis(44,48)	d24	dis(47,49)
d3	dis(20,26)	d14	dis(46,50)	d25	dis(14,23)
d4	dis(22,24)	d15	dis(39,3)	d26	dis(15,27)
d5	dis(29,33)	d16	dis(21,A)	d27	dis(19,23)/2
d6	dis(28,34)	d17	dis(A,25)	d28	dis(27,31)/2
d7	dis(30,32)	d18	hei(A,44)	d29	(wid(19,23)+wid(27,31))/2
d8	dis(39,46)	d19	dis(29,B)	d30	(hei(11,39)+hei(18,39))/2
d9	dis(23,44)	d20	dis(B,33)	d31	(hei(14,39)+hei(15,39))/2
d10	dis(27,48)	d21	hei(B,48)	d32	(hei(44,39)+hei(48,39))/2

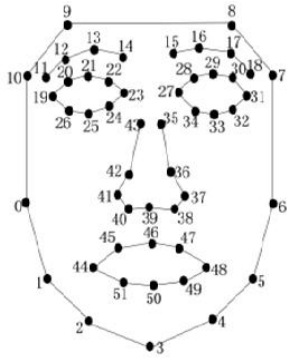


Fig. 3 52 feature points according to FAP parameters

Compared to the method of integrating all the classifiers, we can find that the proposed method is superior to others. Therefore, we can draw a conclusion that dynamically selecting classifier from the candidates is more suitable for emotion recognition than the method of integrating all the candidate classifiers. The results are also consistent to the standpoint of cognitive psychology: the emotions are different from one to another, hence each candidate classifiers is only suitable for a special subset of samples. The most suitable classifier for a new sample should be selected for itself. In some cases, the results are not perfect enough by combining the results of all the classifiers due to the fact that some candidate classifiers may get conflict results for an unseen sample.

In comparison with SARA, we can find that the proposed method is superior to SARA too. Therefore,

we can draw a conclusion that dynamically selecting classifier is more suitable for emotion recognition than selecting fixed classifiers. Although both methods use a single classifier for unseen samples, the proposed method can get a better result since it uses the local ability of unseen samples.

Table 3 Results of the comparative experiments.

Dataset	DEFS		Ensemble all		CEBARKNC		Feature selection	
	n	accuracy	n	accuracy	n	accuracy	n	accuracy
CKACFE	1	81.57	3	78.28	1	73.07	1	76.70
JAFFE	1	69.48	2.75	66.48	1	63.17	1	61.56
CQUPTE	1	89.40	5	87.53	1	78.83	1	89.10
Average	1	80.15	3.58	77.43	1	71.69	1	75.79

4. Conclusions

In this paper, a novel emotion recognition method is proposed based on the ensemble learning. First, a feature selection method is proposed based on RS and domain oriented data driven data mining theory. It can get multiple reductions and candidate classifiers. For each unseen sample, a nearest neighborhood is found and the most accurate classifier is then selected from the candidates. Finally, the selected classifier is used for recognizing the unseen samples. The comparative experiments have validated the proposed method is suitable for emotion recognition.

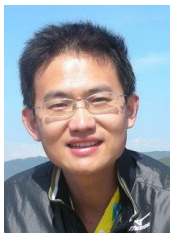
Acknowledgement

This paper is partially supported by Young Teachers Start Research Project of SWJTU (No. 2009Q086), the Fundamental Research Funds for the Central Universities (2009QK17), Chongqing Key Lab of Computer Network and Communication Technology Foundation under Grant No, CY-CNCL-2009-02, Natural Science Foundation of Chongqing University of Posts and Telecommunications under Grant A2009-26, ICST of Chongqing University of Posts and Telecommunication Foundation Grant No, JK-Y-2010002. In particular, the authors wish to thank reviewers for their constructive comments.

References

1. R. W. Picard, Affective Computing: Challenges. International Journal of Human- Computer Studies, 2003, 59(1): 55-64.
2. R. W. Picard, E. Vyzas, and J. Healey. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(10): 1175-1191.
3. T. G. Ditterrich. Machine learning research: four current direction. Artificial Intelligence Magzine, 1997, 4: 97-136.
4. T. G. Ditterrich. Ensemble methods in machine learning. In: Kittler, J., Roli, F., (Eds.), Multiple Classifier Systems. LNCS 2001, 1857: 1-15.
5. A. Tsymbal, M. Pechenizkiy, P. Cunningham. Diversity in search strategies for ensemble feature selection. Information Fusion, 2005, 1: 83-98.
6. L. Breiman. Bagging predictors. Machine Learning. 1996, 2: 123-140.
7. Y. Freund. Boosting a weak algorithm by majority. Information and Computation, 1995, 2: 256-285.
8. T. G. Ditterrich, G. Bakiri. Solving multi-class learning problem via error-correcting output codes. Journal of Artificial Intelligence Research, 1995, 2: 263-286.
9. T. G. Ditterrich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning. 2000, 40(2): 139-157.
10. D. Opitz. Feature selection for ensembles. In: Proceedings of 16th National Conference on Artificial Intelligence, AAAI Press, Florida, 1999: 379-384.
11. G. Y. Wang, Y. Wang. Domain-oriented Data-driven Data Mining: a New Understanding for Data Mining, Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2008, 20(3): 266-271.
12. G. Y. Wang, Introduction to 3DM: Domain-Oriented Data-Driven Data Mining, Proceedings of RSKT2008, 2008: 25-26.
13. Z. Pawlak, Rough sets. International J. Comp. Inform. Science, 1982, 11: 341-356.
14. Z. Pawlak, Rough Classification. International Journal of Man-Machine Studies, 1984, 5: 469-483.
15. S. Ohsuga. Knowledge Discovery as Translation. T. Y. Lin, et al (Eds.): Foundations of Data Mining and Knowledge Discovery, Springer, 2005, 1-19.
16. The Cohn-Kanade AU-Coded Facial Expression

Database. http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html
17. The Japanese Female Facial Expression (JAFFE) Database. http://www.kasrl.org/jaffe.html
18. Chongqing University of Posts and Telecommunications Emotional Database (CQUPTE). http://cs.cqupt.edu.cn/users/904/docs/9317-1.rar.
19. X. Sui, Y. T. Ren, Online Processing of Facial Expression Recognition. Acta Psychologica Sinica, 2007, 39(1): 64-70
20. G.Y. Wang, H. Yu, D.C. Yang. Decision Table Reduction based on Conditional Information Entropy, Chinese Journal of Computers, 2002, 25(7): 759-766.
21. X. Hu, N. Cercone, Learning maximal generalized decision rules via discretization,generalization and rough set feature selection. Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, 1997: 548-556.



Xun Gong received the PhD degree from Southwest Jiaotong University (SWJTU, China) in 2008, and he is presently a lecturer at the Lab of Intelligent Information Processing in SWJTU. His research interests are in computer vision, pattern recognition, image processing and perception computation. His recent research has been concerned with the development of fast and robust methods for realistic 3D face model retrieval and face detection, tracking, and recognition.