ISSN 2320 - 2602
Volume 3, No.5, May 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst07352014.pdf

Facial Emotion Ranking Under Imbalanced Conditions

Yok-Yen Nguwi
James Cook University, Singapore, yokyen.nguwi@jcu.edu.au

ABSTRACT

The aim of emotion recognition is to establish grounds that work for different types of emotions. However, majority of the classifiers have their base from balanced datasets. There are few works that attempts to address how to approach facial emotion recognition under imbalanced condition. This paper discusses the issues related to imbalanced data distribution problem and the common strategy to deal with imbalance datasets. We propose a model capable of handling imbalance emotion datasets well in which other typical classifiers failed to address. The model actively ranks the prototype (i.e. Prototype Ranking) of the facial expression image and adopted a derivation of support vector machines in its selection so that the problem of imbalanced data distribution can be relaxed. Then, we used an Emergent Self-Organizing Map (ESOM) to cluster the ranked features to provide clusters for unsupervised classification. This work progresses by examining the efficiency of the model in evaluating the results to show that the criterion based on prototype ranking achieves good results and performs consistently well over problem domain.

Key words : Imbalanced Datasets, Emotion Recognition, Support Vector Machine, Emergent Self-Organizing Map.

1. INTRODUCTION

A facial expression is formed by contracting or relaxing different facial muscles on human face which results in temporally deformed facial features like wide open mouth, raising eyebrows or etc. Emotions positively affect intelligent functions such as decision making, perception and empathic understanding [1, 2]. Emotion is a state of feeling involving thoughts, physiological changes, and an outward expression. There are five theories which attempt to understand the sequence of processes that we are experiencing when we are feeling a certain type of emotion. They are James-Lange theory [3], Cannon-Bard theory [4], Lazarus theory [5], Schachter-Singer theory [6], and Facial Feedback theory [7]. According to the facial feedback theory [2], emotion is the experience of changes in our facial muscles. In other words, when we smile, we experience pleasure or happiness. When we frown, we experience sadness. It is the changes in our facial muscles that direct our brains and provide the basis for our emotions. As there are many possibilities of muscle configurations in our face, there is seemingly unlimited number of emotions.

Most people are able to interpret the emotions expressed by others all the times, but there are people who lack this ability, such as people diagnosed along the autism spectrum [8]. The first known facial expression analysis was presented by Darwin in 1872 [9]. He presented the universality of human face expressions and the continuity in man and animals. He pointed out that there are specific inborn emotions, which originated in serviceable associated habits. After about a century, Ekman and Friesen [10] postulated six primary emotions that possess each a distinctive content together with a unique facial expression. These prototypic emotional displays are also referred to as basic emotions in many of the later literature. They seem to be universal across human cultures and are namely happiness, sadness, fear, disgust, surprise and anger. They developed the Facial Action Coding System (FACS) for describing facial expressions. It is an appearance-based approach. FACS uses 44 action units [11] for the description of facial actions with regard to their location as well as their intensity. Individual expressions may be modeled by single action units or action unit combinations. FACS codes expression from static pictures. FACS is an anatomically oriented coding system, which is based on the definition of Action Units [11] (AU) of a face causing facial movements. Each Action Unit may correspond to several muscles that generate a certain facial action. Forty-six Action Units were considered responsible for expression control and twelve AUs were assigned for gaze direction and orientation. The Action Unit codes were assigned to the action of muscles specific to certain portion of the face.

Following Ekman, more works evolved during the nineties which include the use of Multi-Layer Perceptron Neural Networks [12], optical flow estimation [13], 2D Potential nets [14], Radial basis function network [15]. Essa and Pentland [16] further extend the FACS and developed a FACS+ model to represent facial expression. More recent works include Ye et al. [17] who use Gabor transformation to form elastic facial graph, Dynamic Bayesian Networks (DBM) with FACS by Ji [18], Xiang et al. [19] utilized fourier transform, fuzzy C means to generate a spatio-temporal model for each expression type. Their recognition rates are in the range of 80~93%. They attempt to recognize the emotions in the general circumstances where the training and testing data do not involve imbalanced inclined data. When a dataset is imbalanced inclined, the number of instances in one class

significantly outnumber the instances from other classes. The imbalanced datasets (IDS) exhibit skewed class distribution in which one class dominates the other. Typical dataset is constructed with equal or close to equal number of instances from each class. Most classifiers perform well in balanced dataset but not imbalanced dataset. According to Rehan [20], classifiers generally do not perform well on imbalanced datasets because they are designed to generalize from sample data and output the simplest hypothesis that best fits the data, based on the principle of Occam's razor. The IDS is observed in many different domains, such as, business, industry, scientific research and many real-world applications like vision recognition [20], bioinformatics [21], credit card fraud detection [22], detection of oil spills [23], medical data [24] and risk management data [25] where certain classes are not as easily available as the other classes. Imbalanced data is not commonly dealt with specifically in conventional classifiers; they do not make special allowance concerning the class imbalance. The IDS usually results in poor performance of standard classification algorithms [23, 26-28] like decision tree, nearest neighbor and naive-bayesian. When the data is imbalanced, it causes problem in proper feature selection and clustering. Traditional machine learning algorithms can be biased towards majority class due to over-prevalence. In the case of decision tree, cases from majority class tend to dominate the tree structure. In k-nearest neighbor, the nearest neighbors are mostly from the majority class. The same goes to naïve-bayesian because majority class is much more probable than minority class. The minority class is usually the class of interest and the errors coming from this class is more important and thus higher penalty errors in cost-sensitive learning.

Another means of emotion recognition is through voice. Voice can provide indications of valence and specific emotions through acoustic properties such as pitch range, rhythm, amplitude, or duration changes [48]. A bored or sad user would typically exhibit slower, lower-pitched speech, with little high frequency energy. A person experiencing fear, anger, or joy would speak faster and louder, with strong high frequency energy [49]. The database of speech recognition generally consists of recorded speech synthesis, prosodic modeling, speech conversion that are voiced out by actors like the work in [50]. Pao et. Al. [44] proposed to solve an interesting problem of Mandarin speech corpus from different emotions using weighted discrete k-nearest neighbor (KNN). The other attempt to solve speech emotion recognition using KNN is by Zhao et. al. who reported strong results in [45].

This work proposes a method that actively ranks the prototype of the face image and anticipates the facial expression based on semi-supervised learning. The facial images are organized into binary problem with varying imbalance ratio. The criteria for prototype selection are derived from Support Vector Machines (SVM) and are based on weight vector sensitivity with respect to variables [29]. Variables are selected using ranking criterion which is a similar way of feature selection aiming at eliminating variables that are less appealing and to enhance the generalization capability of the classifier subsequently used. The use of ranking features reduces the computational times as it does not require the holistic features of the entire face images. In the SVM Emergent Self-Organizing Map (ESOM) [30, 31] is then used for unsupervised classification. Emergent SOM is an extension of Self-Organizing Map (SOM) that allows the emergence of intrinsic structural features of high dimensional data onto a two dimensional map. ESOM is a powerful tool for clustering, visualization, and classification. This work progresses by examining the efficiency of the model in evaluating imbalanced data sets. Experimental results show that the criterion based on weight vector derivative achieves good results and performs consistently well over imbalance data.

In the formalization of SVM, the kernel functions Φ are exploited for the purpose of linear or non-linear mapping input data into high-dimensional feature space $\Phi : R^d$ a H . It is within this novel, the hidden feature space that a simple linear decision surface can be readily designed [32]. Chang et. al. [46] proposed a novel approach to tune weight vector and bias of Support Vector Regression modeling using nested local adiabatic evolution. Xing et. al. [47] combines the auto-correlation wavelet kernel with multiclass least squares support vector machine that enhances generalization ability of SVM. Different kernel function gives origin to different feature spaces and thus different generalization capabilities. This work selects the kernel function by measuring the skewness of the soft margin which can effectively segregate the classes even under the presence of majority data points dominate the minority data points. The idea of soft margin was initiated by Cortes and Vaptik [33] has its root since 1995. Guyon et al [34] further exploit the soft margin to recursively eliminate less discriminative features, recursive feature elimination (RFE), which form the foundation of the prototype ranking in this work. Rakotomamonjy [35] then investigated various selection criteria such as $\|w\|^2$ and criteria based on generalization error bounds in the RFE procedure.

The paper is organized as follows: Section 2 describes some problems stated in the imbalanced data sets and the associated methods to solve such problems. This section also explains the phenomenon of imbalance, and the prototype ranking model to deal with the problem. Section 3 presents the performances of the proposed approach to demonstrate how well of the proposed method to deal with the imbalanced datasets. Finally, conclusion of this paper is drawn in Section 4.

2. IMBALANCE DATA PROTOTYPE RANKING

2.1 Imbalance Phenomenon

Learning from IDS is often reported as a difficult task [36]. It is a phenomenon where the probability estimation of the minority class is much smaller, i.e.

$$P(c = \text{minority}|x) < 0.5 \quad (1)$$

The imbalance ratio is defined as the number of minority over the number of majority,

$$\text{IR} = \frac{\text{Minority Instance}}{\text{Majority Instance}} \quad (2)$$

A simple illustration can be observed from Figure 1 which shows the imbalanced data distribution with IR=0.1. The majority instances are represented by (-), the majority instances are represented by (+). We can see that there are some degrees of class overlapping and the class boundary may not be easily defined.

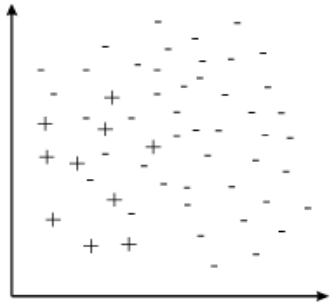


Figure 1. Sample plot of imbalanced data

To a specific dataset $D(x_i, y_i)$, with the distribution $p(Y \mid X)$, for which each $x_i \in X$ the features vector and $y_i \in Y$ is the associated class label. The objective to solve the binary problem is to train a classifier $f : X$ a Y to estimate the probability for each x_i in the testing sample which belong to *Y*. The idea is to enhance the performance by minimizing loss or maximize accuracy. Thus, the model should maximize some objective function, *O*, such that the $E_{(X,Y)}[O(f(X), Y)]$ is maximized. In the presence of high imbalance, this is practically difficult to achieve due to the limitation of the number of instances from minority, so we can only approximate the true function. The most straightforward way to balance the imbalance data is to resample the data. Re-sampling attempts to re-draw the training data distribution to be D^*. By tuning the $S : D$ a D^* can improve the classifier's performance.

We now look at the distributions of data points under radial visualization (also known as Radviz [38]) and the use of sampling to resolve the imbalances. Radviz is a method where the examples are represented by points inside a circle. The visualized attributes correspond to points equidistantly distributed along the circumference of the circle. The most influential attributes wholly decides the position of a point. The data point is placed at the position where the sum of all forces from each attributes equals to zero. Previously we discussed that sampling is thus far the most popular approach in solving imbalanced data problem. However, as shown in Figure 2 are the visualizations of imbalanced and balanced datasets for emotion data. The datasets are originally imbalanced with much larger numbers of negative instances. We randomly down sampled the positive instances to achieve a balance ratio between the two. Radviz also suggests that the down sampling does not make the data separable under the presence of full features. The dataset is arranged into m-binary problem with m=4. The 174 data points from balanced happy emotions dataset appear to be centrally clustered together for both the classes. The sad emotion dataset consists of 207 data points initially as shown in Figure 3(c) which is dominated by class 1. After applying the sampling, the data distribution remains the same with a balanced IR. The same goes with the fear and surprise emotion data. Radviz showed that the intrinsic properties of imbalanced datasets remain the same despite randomly down-sample it to have a balanced number of samples as the data distributions are inseparable.

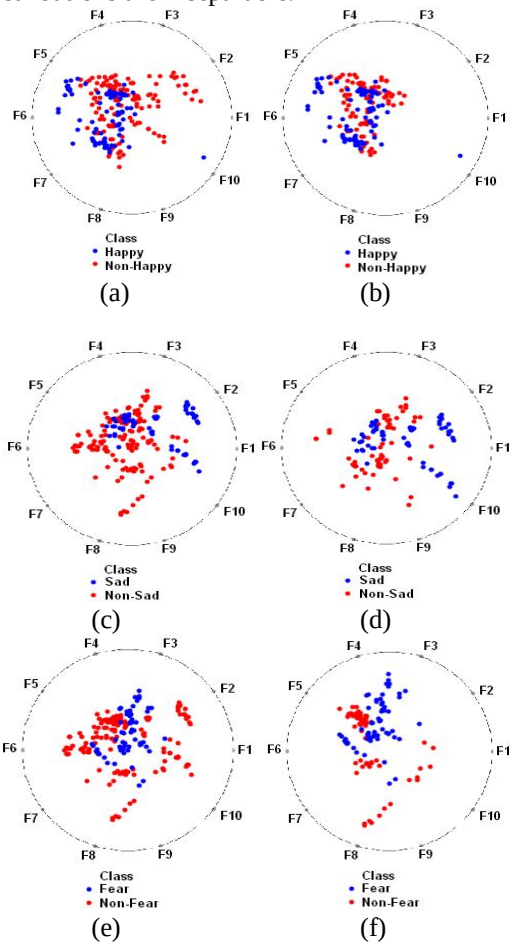


Figure 2. Multi-dimensional visualization of the distribution of balanced and imbalanced datasets; (a) Imbalanced happy data; (b) Balanced happy data; (c) Imbalanced sad data; (d) Balanced sad data; (e) Imbalanced fear data; (f) Balanced fear data.

2.2 Prototype Ranking

The process started with Recursive Partitioning of the Majority Class (REPMAC) approach to translate the imbalanced problem into a set of balanced problem. The REPMAC [39] first apply a clustering method to the majority, in order to obtain two clusters. We analyze each of the clusters to check if they meet any of the set of stopping criteria. If they do not, we apply again the clustering method (thus creating a recursive process). When one of the clusters meets the criteria, we fit the classifier to the resulting dataset.

Algorithm 1: The REPMAC Method

Inputs:

D^+ : The majority class dataset

D^- : The minority class dataset

$Cl()$: Clustering method

$DF()$: Decision function

Function REPMAC(D^+, D^-, Cl, DF):

1. Apply Cl to D^+ to create D_1^+ and D_2^+
2. For *i*=1 to 2:

   IF Stopping-Criteria $\left(D_i^+, D^-\right)$ is met THEN

   Build a classifier $DF(D_i^+, D^-)$

   ELSE

   CALL $REPMAC(D_i^+, D^-, Cl, DF)$:

The next step in prototype ranking is SVM training as shown in Figure 3. The prototypes are batch trained together with the criterion. The remaining values of surviving features continued to be recursively ranked till all criterions are met and all rankings are finalized. The prototype rankings are derived from Support Vector Machines and are based on weight vector sensitivity with respect to a prototype. It was initially proposed by Guyon *et al.* [40] for selecting genes that is relevant for a cancer classification problem. The squared coefficients $w_j^2 (j = 1,..., p)$ of the weight vector w are employed as feature ranking criteria. The prototypes in this context represent the grey intensity of the pixel. Prototypes are selected using ranking criterion to rank variables. The ranking criteria w_j^2 for all features are computed, and the prototype with the smallest ranking criterion is discarded. This is an extension to bounds on L error, margin bound and other bounds of the generalization error. The criterion being investigated is C_t which is either weight vector $\|w\|^2$, the radius/margin bound $R^2\|w\|^2$ or the span estimate. It gives either an estimation of the generalization performance or an estimation of the dataset separability. It was initially proposed by Guyon *et al.* [40] for selecting genes that is relevant for a cancer classification problem. The goal is to find a subset of size r among d features where $r < d$ which maximizes the performance of classifier [41]. The method is based on backward sequential selection. The features are removed one at a time until r features. The criteria for selection are derived from SVM and are based on weight vector sensitivity with respect to a variable [40]. Variables are selected using ranking criterion to rank variables. This is an extension to bounds on L error, margin bound and other bounds of the generalization error [41]. The criteria being investigated is C_t which is either weight vector $\|\mathbf{w}\|^2$, the radius/margin bound $R^2\|\mathbf{w}\|^2$ or the span estimate. It gives either an estimation of the generalization performance or an estimation of the dataset separability. The removed variable is the one whose removal minimizes the variation of $\|\mathbf{w}\|^2$. Hence, the ranking criterion R_c for a given variable *i* is:

$$\left|\|\mathbf{w}\|^2 - \|\mathbf{w}^{(i)}\|^2\right| = \frac{1}{2}\left|\sum_{k,j}\alpha_k^*\alpha_j^* y_k y_j K(\mathbf{x}_k,\mathbf{x}_j) - \sum_{k,j}\alpha_k^{*(i)}\alpha_j^{*(i)} y_k y_j K^i(\mathbf{x}_k,\mathbf{x}_j)\right| \quad (3)$$

where $K^{(i)}$ is the Gram matrix of the training data when variable *i* is removed ($K_{k,j}^i = \left\langle \Phi(\mathbf{x}_k^{(i)}), \Phi(\mathbf{x}_j^{(i)}) \right\rangle$) and $\alpha_k^{*(i)}$ is the solution of quadratic optimization problem. From the above equation, the removed variable is the one which has the least influence on the weight vector norm.

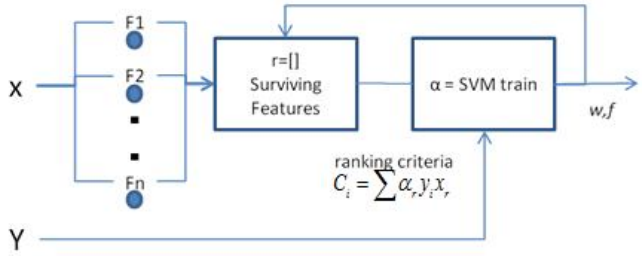


Figure 3. Overview of prototype ranking

This approach of criteria ranking derived from SVM is different from the popularly used soft margin SVM. In traditional SVM where a given set of labeled instances $\mathbf{X}_{train} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ and a kernel function k, SVM finds the optimal α_t for each x_i to maximize the margin γ between the hyperplane and the closest instances to it. The prediction for a new sample x is made through:

$$sign\left(f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \quad (4)$$

where *b* is the threshold. One norm soft-margin SVM minimizes the primal Lagranian:

$$L_p = \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i\xi_i \quad (5)$$

where $\alpha_i \geq 0$ and $r_i \geq 0$. The penalty constant C represents the trade-off between the empirical error ξ and the margin. Two approaches [41] are proposed for each criterion:

- *Zero-order method*: The criterion C_t is directly used for variable ranking, and it identifies the variable that produces the smallest value of C_t when removed. The ranking criterion becomes $R_c(i) = C_t^{(i)}$ with $C_t^{(i)}$ being the criterion value when variable *i* has been removed.
- *First-order method*: This uses the derivatives of the criterion C_t with regards to a variable. This approach differs from the zero-order method because a variable is ranked according to its influence on the criterion which is measured with the absolute value of the derivative.

The zero-order criteria based on bounds have been used for feature selection associated with different search space algorithm whereas the first-order ones are rather new for the purpose of feature selection. In the zero-order method, one suppresses the feature whose removal minimizes the criterion whereas in the first order methods, one removes the variable to which the criterion is less sensitive. For instance, in the zero-order $\|\mathbf{w}\|^2$ case, the ranking term is:

$$R_c(i) = \left\|\mathbf{w}^{(i)}\right\|^2 = \sum_{k,j} \alpha_k^{*(i)} \alpha_j^{*(i)} y_k y_j K^i\left(\mathbf{x}_k, \mathbf{x}_j\right) \tag{6}$$

where $K^{(i)}$ is the Gram matrix of the training data when variable *i* is removed.

In the first order case, the sensitivity of a given criterion with respect to a variable is measured. A possible approach is to introduce a virtual scaling factor and to compute the gradient of a criterion with respect to that scaling factor ρ. The latter acts as a component wise multiplicative term whose value is 1 on the input variables and thus $k\left(\mathbf{x}, \mathbf{x}'\right)$ becomes:

$$k\left(\rho \cdot \mathbf{x}, \rho \cdot \mathbf{x}'\right), \tag{7}$$

where "·" denotes the component wise vector product. Consequently, one obtains the following derivatives for a Gaussian Kernel.

$$k\left(\rho \cdot \mathbf{x}, \rho \cdot \mathbf{x}'\right) = \exp\left(-\frac{\left\|\rho \cdot \mathbf{x} - \rho \cdot \mathbf{x}'\right\|}{2\sigma^2}\right):$$

$$\frac{\partial k}{\partial \rho_i} = -\frac{1}{\sigma^2}\left(\rho_i x_i - \rho_i x_i'\right)^2 k\left(\mathbf{x}, \mathbf{x}'\right) = -\frac{1}{\sigma^2}\left(x_i - x_i'\right)^2 k\left(\mathbf{x}, \mathbf{x}'\right), \tag{8}$$

where we used $\rho_i = 1$. Then one needs to evaluate the gradient of the bounds with regards to a variable ρ_i and for a given criterion *C*, the ranking term would be:

$$R_c(i) = \left|\frac{\partial C(\alpha, b)}{\partial \rho_i}\right| = \left|\nabla \left\|\mathbf{w}^{(i)}\right\|^2\right|, \tag{9}$$

The problem of searching the best *r* variables is solved by means of greedy algorithm based on backward selection. A backward sequential selection is used because of its lower computational complexity compared to randomized or exponential algorithms and its optimality in the subset selection problem. The algorithm starts with all features and repeatedly removes a feature until *r* features are left for all variables have been ranked.

Algorithm 2: SVM based Ranking for Variable Selection

Given Rank=[] ; Var=[1,...,N]

Repeat the following until Var is not empty:

1. Train a SVM classifier with all the training data and variable Var.
2. For all Var, evaluate ranking criterion $R_c(i)$ of variable *i*
3. $\text{Best} = \arg\min R_c$
4. Rank the variable that minimizes R_c :

   Rank=[bestRank]
5. Remove the variable that minimizes R_c from the selected variables set Var=[1, …, best-1, best+1, …N]

2.3 Semi-supervised Classification

After finding the top ranked variables, we then use the features for emergent self-organizing mapping. The Emergent Self-Organizing Map is a non-linear projection technique using neurons arranged on a map. There are mainly two types of ESOM grid structures in use: hexgrid (honeycomb like) and quadgrid (trellis like) maps.

ESOM forms a low dimensional grid of high dimensional prototype vectors. The density of data in the vicinity of the models associated with the map neurons, and the distances between the models, are taken into account for better visualization. An ESOM map consists of a U-Map (from U-Matrix), a P-Map (from P-Matrix) and a U*-Map (which combines the U and P map). The three maps show the floor space layout for a landscape like visualization for distance and density structure of the high dimensional data space. Structures emerge on top of the map by the cooperation of many neurons. These emerging structures are the main concept of ESOM. It can be used to achieve visualization, clustering, and classification. This can be obtained by the following.

Let $m: D \to M$ be a mapping from a high dimensional data space $D \subset \Re^n$ onto a finite set of positions $M = \{n_1, \text{K}, n_k\} \subset \Re^2$ arranged on a grid. Each position has its two dimensional coordinates and a weight vector $\mathbf{W} = \{w_1, \text{K}, w_k\}$ which is the image of a Voronoi

region in *D*: the data set $E = \{x_1, \mathrm{K}, x_d\}$ with $x_i \in D$ is mapped to a position in *M* such that a data point x_i is mapped to its best-match $bm(x_i) = n_b \in M$ with $d(x, w_b) \le d(x, w_j);\ \forall w_j \in \mathbf{W}$, where *d* is the distance on the data set. The set of immediate neighbors of a position n_i on the grid is denoted by $N(i)$.

The clustering of ESOM is based on the U*C clustering algorithm described by [31]. Consider a data point *x* at the surface of a cluster *C*, with a best match of $n_i = bm(x)$. The weight vectors of its neighbors $N(i)$ are either within the cluster, in a different cluster or interpolate between clusters. Assume that the inter cluster distances are locally larger than the local within-cluster distances, then the U-heights in $N(i)$ will be large in such directions which point away from the cluster *C*. Thus, a so-called immersive movement will perform to lead away from cluster borders. This immersive movement is performed which starts from a grid position, keeps decreasing the U-matrix value by moving to the neighbor with the smallest value, then keeps increasing the P-matrix value by moving to the neighbor with the largest value. The details of this clustering algorithm can be referred to [31]. The SVESOM proceeds as follows:

Algorithm 3: Unsupervised clustering

Given U-Matrix, P-Matrix, U*-Matrix, I={}.

Immersion:

For all positions *n* of the grid:

1. From position *n* follow a descending movement on the U-Matrix until the lowest distance value is reached in position *u*.
2. From position *u* follow an ascending movement on the P-Matrix until the highest density value is reached in position *p*.
3. $I = I \cup \{p\}$; $\mathrm{Immersion}(n) = p$.

Cluster assignment:

1. Calculate the watersheds for the U*-Matrix using the algorithm in [30].
2. Partition *I* using these watersheds into clusters $C_1 \mathrm{K}\ C_c$.
3. Assign a data point *x* to a cluster C_j if $\mathrm{Immersion}(bm(x)) \in C_j$.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were carried out using three stratified cross-validation training and testing sets. Two third of the dataset constitute the training sets and one third for testing. We have adopted datasets from two sources: Carnegie Mellon University(CMU)'s Cohn-Kanade AU-coded Facial Expression database (The face research group [42]). CMU database consists of images of approximately 100 subjects. Facial images are of size 640x490 pixels, 8-bit precision grayscale in png format. Subjects were 100 university students enrolled in introductory psychology class. Age ranges from 18 to 30. Sixty-five percent were female, 15 percent were African-American, and three percent were Asian or Latino. Subjects were instructed by experimenter to perform a series of facial displays. Subjects began each display from a neutral face. Before performing each display, the experimenter described and modeled the desired display.

Accuracy may not be an appropriate measure to evaluate the performance of classifiers conducting for imbalanced datasets. A classifier may be able to obtain very high accuracy by classifying all instances to majority class but outnumbers the minority class. Therefore, the use of accuracy and mean square error rate are inappropriate for imbalanced dataset like emotion. So we adopted the use of other kinds of evaluation metrics to measure the classifiers' capability to differentiate the two-class problem under the case of imbalanced data. They are namely F-measure and Geometric Means Measure (GMM). These performance measures are independent to prior probabilities. F-measure is a metric derived from recall and precision. Some variants using different weighting would make them equal weighted as we consider false positive and false negative equal likely to occur. According to the evaluations by Barandela *et al.* [43], geometric means measure (GMM) is a more appropriate metric to evaluate the classifier performance on IDS. Both Receiver Operating Characteristic (ROC) curve and GMM are good indicators as they try to maximize the accuracy on each of the two classes while keeping these accuracies balanced. The GMM is defined as the square root of the product of accuracy on positive samples and negative samples.

The data are organized into binary problem. We attempt to recognize the four most basic emotions namely happy, sad, fear, and surprise. To recognize happy, the other emotions are combined together to form non-happy class. The sad, fear, and surprise emotions are organized in the similar manner. Table 1 tabulates the results obtained from the experiments. The IR denotes the imbalance ratio, Gn denotes generalization, Rc denotes recalling, followed by F-measure, Geometric means for test set, and Geometric means for training set. The first table shows the result of PR with balanced data. The data ratios are about the same for both classes. The IR are designed in the range of 0.9~1. The balanced data usually do not pose a problem, so the focus of our experiments is on imbalanced emotions. Table 1 shows the emotions with varying IR from 0.27 to 0.6. The results show that the F-measures are consistently above 90%. We benchmark the results with the use of Principle Components Analysis (PCA), Independent Component Analysis (ICA), GINI and T-Test. They are well-known dimension reduction techniques. Table 2 shows that the results vary between 61% and 99%. Even PCA pose inconsistent results under imbalance conditions. The results for ICA are even weaker, between 61~83% for F-measure. The classifier in used is emergent self-organizing map. GINI index works by traversing all the possible segmentation method for each attribute and provides for a GINI index. The feature with the minimum Gini index is selected. The GINI offers good result which is quite comparable to our PR approach with fear emotion being the most discriminative emotion, the results vary from 0.8~0.93. T-Test is a statistical approach to select

features base on the means of two classes. T-test result is slightly weaker than GINI.

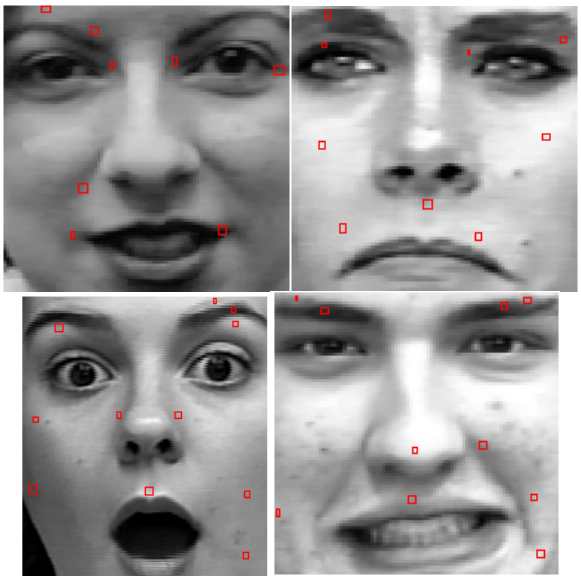
Table 1. Performance of PR with imbalanced data

	IR	Gn	Rc	F-Measure	GMts	GMtr
Happy	0.60	89.1%	96.9%	92.8%	88.9%	97.4%
Sad	0.44	90.9%	98.1%	94.4%	90.1%	97.7%
Fear	0.44	97.7%	100.0%	98.9%	98.3%	100.0%
Surprise	0.27	94.4%	100.0%	97.1%	91.0%	100.0%

Table 2. Performance Evaluation of Different Feature Selection Methods

Method	Happy		Sad		Fear		Surprise	
	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure
PR	**0.89**	**0.93**	**0.91**	**0.94**	**0.98**	**0.99**	**0.94**	**0.97**
PCA	0.7	0.82	0.8	0.89	0.77	0.7	0.65	0.68
ICA	0.8	0.8	0.73	0.83	0.59	0.62	0.61	0.65
GINI	0.87	0.92	0.8	0.87	0.91	0.93	0.86	0.91
T-Test	0.75	0.83	0.68	0.8	0.8	0.89	0.94	0.97

Figure 4 shows the images used during the experiments, together with ten important features that are obtained from the SVM-trainer. The features lie on points that significantly differentiate the facial expression. Take happy for example, a happy face is generally made up of raised eyebrows, and lips corner Figure 4(a). Figure 4(b) shows a sad face with downwards lips corner, the PR is able to select such significant points that contributes to good performance in classifier. As for the case of surprise expression in Figure 4(c), the cheeks are usually stretched longer, and the results of PR are focus on selecting pixels in these areas. Lastly is the fear emotion as shown in Figure 4(d), which is slightly more challenging. The PR selects pixels evenly distributed across the face image, from upper eyelids, to nose, cheeks, and lips.



(a) (b) (c) (d)

Figure 4. Important features from SVM-ranking

4. CONCLUSION

This paper presents a new approach of learning from imbalanced datasets through combining a ranking method with emergent self-organizing approach. This work first employs the use of a derivation of support vector for variable selection and the criteria for selection are derived from Support Vector Machines (SVMs). It is based on weight vector sensitivity with respect to a variable. After that, the Emergent Self-Organizing Map (ESOM) is employed for the feature mapping which has the capability to visualize the distance structure as well as the density structure of high dimensional data sets. It is suitable to detect non-trivial cluster structures which are benefit to the performance of the unsupervised classification. In the experiments with different facial expressions, our proposed method is shown to be effective and better than several other methods like Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The aim of this work is to resolve emotion recognition problem even under imbalanced condition. This is an important area, as a classifier cannot be termed effective if it is only able to segregate balanced data.

REFERENCES

1. Bechara, A., H. Damasio, and A.R. Damasio, Emotion decision making and the orbitofrontal cortex. Cerebral Cortex, 2000 10(3): p. 295–307.
2. Isen, A.M., Positive affect and decision making. Handbook of emotions. 2000, New York: Guilford Press.
3. Cook, H.D., The James-Lange theory of the emotions and the sensationalistic analysis of thinking. Psychological Bulletin, 1911. 8(3): p. 101-106.
4. Newman, I., Cannon's theory of emotion, and an alternative thalamic theory. Journal of Abnormal and Social Psychology, 1936. 31(3): p. 253-259.
5. Lazarus, R.S., Thoughts on the relations between emotion and cognition. American Psychologist, 1982. 37(9): p. 1019-1024.
6. Reisenzein, R., The Schachter theory of emotion: Two decades later. Psychological Bulletin, 1983. 94(2): p. 239-264.
7. Buck, R., Nonverbal behavior and the theory of emotion: The facial feedback hypothesis. Journal of Personality and Social Psychology, 1980. 38(5): p. 811-824.
8. Baron-Cohen, S., Mindblindness: An essay on autism and theory of mind. 1995 Cambridge: MIT Press.
9. Darwin, C., The Expression of the Emotions in Man and Animals. 1872: J. Murray, London.
10. Ekman, P. and W.V. Friesen, Constants across cultures in the face and emotion. J. Personality Social Psychol, 1971. 17(2): p. 124-129.
11. Bauer, H.-U. and K.R. Pawelzik, Quantifying the neighborhood preservation of self-organizing feature maps. IEEE Transactions on Neural Networks, 1992. 3(4): p. 570-579.
12. Cottrell, G.W. and M.K. Fleming. Categorisation of Faces Using Unsupervised Feature Extraction. in Int'l Conf. Neural Networks. 1990. San Diego.
13. Mase, K. and A. Pentland, Recognition of facial expression from optical flow. Institute of Electronics

Information and Communication Engineers (IEICE) Trans., 1991. 74(10): p. 3474-3483.
14. Matsuno, K., C.-W. Iee, and S. Tsuji. Recognition of Human Facial Expressions Without Feature Extraction. in ECCV. 1994.
15. Rosenblum, M., Y. Yacoob, and L.S. Davis, Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture. IEEE Transactions on Neural Networks, 1996. 7(5): p. 1121-1138.
16. Essa, I.A. and A.P. Pentland, Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. 19(7): p. 757-763./'////
17. Ye, J., Y. Zhan, and S. Song. Facial expression features extraction based on//l,, Gabor wavelet transformation. in IEEE International Conference on System, Man and Cybernatics. 2004.
18. Ji, Y.Z.Q., Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. 27(5): p. 699-714.
19. T. Xiang, M.K.H.L.a.S.Y.C., Expression recognition using fuzzy spatio-temporal modeling. Pattern Recognition, 2007. 41(1): p. 204-216.
20. Akbani, R., S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. in Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). 2004.
21. Chawla, N.V., N. Japkowicz, and e. A. Kolcz. Special Issue on Learning from Imbalanced Data Sets. in Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets. 2003.
22. Chawla, N., N. Japkowicz, and Z.H. Zhou, Data Mining When Classes are Imbalanced and Errors Have Costs, in PAKDD'2009 Workshop. 2009: Thailand.
23. Kubat, M., R. Holte, and S. Matwin, Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 1998: p. 195-215.
24. Provost, F. and T. Fawcett, Robust classification for imprecise environments. Machine Learning, 2001. 42: p. 203-231.
25. Drummond, C. and R.C. Holte. Explicitly representing expected cost: An alternative to ROC representation. in Proceeding of the Sixth ACM SIGKDD (Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery and Data Mining . 2000.
26. Maloof, M. Learning when data sets are imbalanced and when costs are unequal and unknown. in Proceedings of the ICML(International Conference on Machine Learning) Workshop on Learning from Imbalanced Data Sets. 2003.
27. Guo-ping, L., Y. Li-xiu, and Y. Jie, Solving the Problem of Imbalanced Dataset in the Pre-diction of Membrane Protein Types Based on Weighted SVM. Journal of Shanghai Jiaotong University, 2005: p. 1676-1684.
28. Chan, P. and S. Stolfo. Toward scalable learning with nonuniform class and cost distributions: A case study in credit card fraud detection. in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining,. 1998.
29. Rakotomamonjy, A., Variable selection using svm based criteria. The Journal of Machine Learning Research 2003: p. 1357-1370.
30. Ultsch, A. and M. F., ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. 2005, University of Marburg Dept. of Mathematics and Computer Science: Germany.
31. Ultsch, A., Clustering with SOM: U*C, in WSOM. 2005. p. 75-82.
32. Lima, C.A.M., A.L.V. Coelho, and F.J. VonZuben, Pattern classification with mixtures of weighted least-squares support vector machine experts. Neural Computing and Applications, 2009. 18(7): p. 843-860.
33. Cortes, C. and V. Vapnik, Support-vector networks. Machine Learning, 1995. 20(3): p. 273-297.
34. Guyon, I., et al., Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 2002. 46(1): p. 389-422.
35. Rakotomamonjy, A., Variable selection using SVM based criteria. J. Mach. Learning, 2003: p. 1357–1370.
36. Batista, G., R.C. Prati, and M.C. Monard, A study of the bahavior of several methods for balancing machine learning training data. SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations, 2004. 6(20-29).
37. Cieslak, D.A. and N.V. Chawla. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. in Proceedings - IEEE International Conference on Data Mining, ICDM. 2008.
38. Hoffman, P., et al. DNA visual and analytic data mining. in Visualization '97., Proceedings. 1997.
39. Ahumada, H., et al. REPMAC: A new hybrid approach to highly imbalanced classification problems. in Proceedings - 8th International Conference on Hybrid Intelligent Systems, HIS 2008. 2008.
40. Guyon, I., et al., Gene Selection for Cancer Classification using Support Vector Machines Machine Learning, 2002. 46(1-3): p. 389-422.
41. Rakotomamonjy, A., Variable selection using svm based criteria. The Journal of Machine Learning Research, 2003: p. 1357-1370.
42. The_face_research_group, CMU Image Data Base http://vasc.ri.cmu.edu/idb/html/face/.
43. Barandela, R., et al., Strategies for learning in class imbalance problems. Pattern Recognition, 2003. 36(3): p. 849–851.
44. Tsanglong Pao, Yute Chen and Junheng Yeh, Emotion Recognition and Evaluation from Mandarin Speech Signals, International Journal of Innovative Computing,

Information and Control, vol.4, no.7, pp.1695-1710, 2008.

45. Lasheng Zhao, Xiaopeng Wei, Qiang Zhang and Rui Liu, Speech Emotion Recognition Based on a Modified k-Nearest Neighbor Algorithm, ICIC Express Letters, vol.4, no.4, pp.1311-1318, 2010.
46. Bao Rong Chang and Hsiu Fen Tsai, Training Support Vector Regression by Quantum-neuron-based Hopfield Neural Net with Nested Local Adiabatic Evolution, International Journal of Innovative Computing, Information and Control, vol.5, no.4, pp.1013-1026, 2009.
47. Y. Xing, X. Wu and Z. Xu, Multiclass Least Squares Auto-correlation Wavelet Support Vector Machines, ICIC Express Letters, vol.2, no.4, pp.345-350, 2008.
48. Ball G. and J. Breese, 2000. Emotion and personality in a conversational agent. In S. Prevost J. Cassell, J. Sullivan and E. Churchill editors. Embodied Conversational Characters. Cambridge, MA: MIT Press.
49. Picard, R., 1997. Affective Computing. The MIT Press, Cambridge, MA.
50. Syaheerah L. Lutfi, J. M. Montero, R. Barra-Chicote, J. M. Lucas-Cuesta, Ascensión Gallardo-Antolín: Expressive Speech Identifications based on Hidden Markov Model. HEALTHINF 2009: 488-494