International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 29, No. 3 (2015) 1551002 (17 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0218001415510027

World Scientific
www.worldscientific.com

Using Skew for Classification

Mark J. DeBonis

Department of Mathematics, Manhattan College
4513 Manhattan College Parkway
Riverdale, NY 10471, USA
mark.debonis@manhattan.edu

Received 21 March 2014
Accepted 12 February 2015
Published 23 March 2015

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

One classic example of a binary classifier is one which employs the mean and standard deviation of the data set as a mechanism for classification. Indeed, principle component analysis has played a major role in this effort. In this paper, we propose that one should also include skew in order to make this method of classification a little more precise. One needs a simple probability distribution function which can be easily fit to a data set and use this pdf to create a classifier with improved error rates and comparable to other classifiers.

Keywords: Machine learning; skew; classification; modeling.

1. Introduction

Principle component analysis (PCA) is a integral and wide-spread technique used in classification. One need only search for PCA in scholarly journals to realize the huge number of citings (as an example, see Ref. 12). Mahalanobis distance goes hand-in-hand with PCA and oftentimes is used as a criterion to determine to which class a particular data point belongs. In view of the fact that PCA finds the directions of maximal variance for projections of the data onto lines passing through its mean, we see that PCA relies heavily on the mean and variance of the data in order to make decisions regarding classification. Indeed, PCA is using these two statistics in order to model the shape of each class.

In this paper, we propose adding one more statistics, namely skew, in order to obtain a more accurate model of each class and thus obtain a more accurate method of classification. With actual data sets, oftentimes the classes are not symmetric in each component and exhibit significant skew. Indeed, many examples of skewed data can be found in economics, finance and biostatistics.[1,2] PCA in a sense is assuming that the data is Gaussian and this in and of itself may be a debilitating assumption. We have found that by introducing skew into our classification the error rate reduces

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

significantly and as a classifier is comparable in accuracy with known ones. Of course, there are other researchers who have considered using non-Gaussian methods for modeling data sets (see for instance Refs. 7 and 11), but what we present here is a novel way of introducing a metric whose distance takes into account the skew of the data.

Therefore, instead of modeling the data with normal distributions we sought a two-tailed distribution which can achieve any amount of skew.

2. Skew Probability Distribution

In Ref. 4, the authors put forth a simple and intuitive way to take an existing unimodal symmetric (around zero) probability distribution and make an associated skew distribution. Indeed, it is easy to describe: given a unimodal symmetric (around zero) distribution $f(x)$ and a scalar $k > 0$, define

$$g(x) = \frac{2k}{k^2+1}\begin{cases} f(kx), & x < 0, \\ f(x/k), & x \geq 0. \end{cases}$$

It is pointed out in Ref. 4, and it is easy to verify, that in general the skewness of $g(x)$ is bounded by the skew of one tail of the symmetric distribution $f(x)$. In the case of the normal distribution, the skewness of $g(x)$ is roughly bounded by 1. In the case of the Laplace distribution, the skewness of $g(x)$ is bounded by 2. This is obviously a serious limitation if we wish to model data with skew higher than 2. Of course, others have explored the construction of skew probability distributions,[5] but we sought a distribution which has the freedom to take on any value with respect to skew and one which is easy to compute and fit to a data set.

We chose a special case of the generalized Cauchy distribution (see Ref. 10 where it was first introduced).

$$f(x) = \frac{a-1}{2(1+|x|)^a} = \frac{a-1}{2}(1+|x|)^{-a}, \quad \text{for } a > 0.$$

The skewed generalized Cauchy distribution becomes

$$g(x) = \frac{(a-1)k}{1+k^2}\begin{cases} (1+k|x|)^{-a}, & \text{for } x < 0, \\ \left(1+\dfrac{|x|}{k}\right)^{-a}, & \text{for } x \geq 0. \end{cases}$$

In Appendix A, we prove the following relevant results regarding $g(x)$:

(1) The skew can be made as large as we like by letting a tend to 4 from the right in the formula above.

(2) When k is known, the maximum likelihood estimator

$$\hat{a} = \frac{N}{\sum_{x_i<0} \ln(1+k|x_i|) + \sum_{x_i \geq 0} \ln\left(1+\frac{|x_i|}{k}\right)} + 1.$$

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

(3) When both a and k are unknown, the maximum likelihood estimator $\widehat{k}$ is the unique point of intersection of the curves

$$y_3(k) = \frac{\frac{N(k^2-1)}{k^2+1}}{\sum_{x_i<0} \frac{kx_i}{1-kx_i} + \sum_{x_i \geq 0} \frac{x_i}{k+x_i}} \quad \text{and}$$

$$y_4(k) = \frac{N}{\sum_{x_i<0} \ln(1+k|x_i|) + \sum_{x_i \geq 0} \ln\left(1+\frac{|x_i|}{k}\right)} + 1.$$

The value of $\hat{a}$ is then obtained as in part 2.

3. One-Sided Standard Deviations

With a notion of skew distance in mind we define the following one-tailed standard deviation for a probability distribution:

Definition 1: Let $p(x)$ be a probability distribution. The positive standard deviation, denoted σ_+, of $p(x)$ is defined to be the standard deviation of the probability distribution $p^+(x) = Cp(|x|)$ for appropriate constant C. The negative standard deviation, denoted σ_-, is defined in a similar manner for $p^-(x) = Cp(-|x|)$.

Now consider the general skew distribution function $g(x)$ defined in the previous section. Consider the two unimodal symmetric (around zero) functions $g^+(x) = g(|x|)$ and $g^-(x) = g(-|x|)$.

Of course, g^+ and g^- are not probability density functions (pdfs) (except when $k = 1$). Let p^+ and p^- be the associated probability functions, respectively. We compute p^+:

$$\int_{-\infty}^{\infty} g^+(x)dx = 2\int_0^{\infty} g(x)dx = \frac{4k}{k^2+1}\int_0^{\infty} f(x/k)dx = \frac{2k^2}{k^2+1}.$$

Hence,

$$p^+(x) = \left(\frac{k^2+1}{2k^2}\right)g^+(x) = \left(\frac{k^2+1}{2k^2}\right)\left(\frac{2k}{k^2+1}\right)f(|x|/k) = \frac{1}{k}f(|x|/k).$$

Similarly, it can be shown that $p^-(x) = kf(-k|x|)$. Let X, X_+ and X_- be the c.r.v.'s (continuous random variables) associated with the pdf's f, p^+ and p^-, respectively. One can easily compute that

$$\sigma_{X_+} = k\sigma_X \quad \text{and} \quad \sigma_{X_-} = \sigma_X/k.$$

4. Skew Error Ellipse

Let us first consider the case of a one-dimensional confidence interval. Suppose we have a collection of data with mean $\mu = 0$, standard deviation σ and pdf f which is symmetric about its mean.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

Theorem 1. *Let $0 \leq \alpha \leq 1$ and $c \in \mathbb{R}$. If $[-c\sigma \; c\sigma]$ is the $\alpha \times 100\%$ confidence interval for a unimodal symmetric around zero pdf, then $[-c\sigma_- \; c\sigma_+]$ is the $\alpha \times 100\%$ confidence interval for the associated skew distribution g.*

Proof.

$$\begin{aligned}
\int_{-c\sigma_-}^{c\sigma_+} g(x)dx &= \frac{2k}{k^2+1}\left(\int_{-c\sigma/k}^{0} f(kx)dx + \int_{0}^{ck\sigma} f(x/k)dx\right)\\
&= \frac{2}{k^2+1}\left(\int_{-c\sigma}^{0} f(u)du + k^2\int_{0}^{c\sigma} f(u)du\right)\\
&= \frac{2}{k^2+1}\left(\frac{1}{2}\alpha + k^2\frac{1}{2}\alpha\right) = \alpha.
\end{aligned}$$

□

Let us consider the two-dimensional case in order to motivate the general result in higher dimensions. As is typically done, we first approximate the pedal curve with an ellipse. The axes of these ellipses are derived from the eigenvectors and eigenvalues of the covariance matrix for the data. To extend our result from skew confidence intervals to skew ellipses, consider the projection of the data onto each of the eigenvalues and in each of these single dimensions proceed as we did in the one-dimensional case. First, let us define more precisely a skew error ellipse. In the first quadrant it is an ellipse with axes σ_{X_+} and σ_{Y_+}. In the second, σ_{X_-} and σ_{Y_+}. In the third σ_{X_-} and σ_{Y_-}. In the fourth σ_{X_+} and σ_{Y_-}.

The potential benefits that the skew error ellipse can have over the standard error ellipse are the following: Fix a probability $0 \leq \alpha \leq 1$ and consider the corresponding standard error ellipse and skew error ellipse.

(1) The skew error ellipse can conform better to the shape of the data than the standard error ellipse.
(2) The skew error ellipse can cover a comparable area to the standard error ellipse (shown below for certain distributions).

In Fig. 1, we generated random data from a generalized Cauchy distribution and fit it with both a standard and skew error ellipse.

The following is a generalization of Theorem 1:

Theorem 2. *Assuming the random variables X and Y are independent, consider the $\alpha \times 100\%$ confidence ellipse with axes $c\sigma_X$ and $c\sigma_Y$, for a joint probability density function which is a product of unimodal symmetric around zero pdfs f_X and f_Y. Then the corresponding skew ellipse with axes $c\sigma_{X_-}$, $c\sigma_{X_+}$, $c\sigma_{Y_-}$ and $c\sigma_{Y_+}$ is the $\alpha \times 100\%$ confidence ellipse for the joint pdf which is a product of the associated skew distribution functions g_X and g_Y.*

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

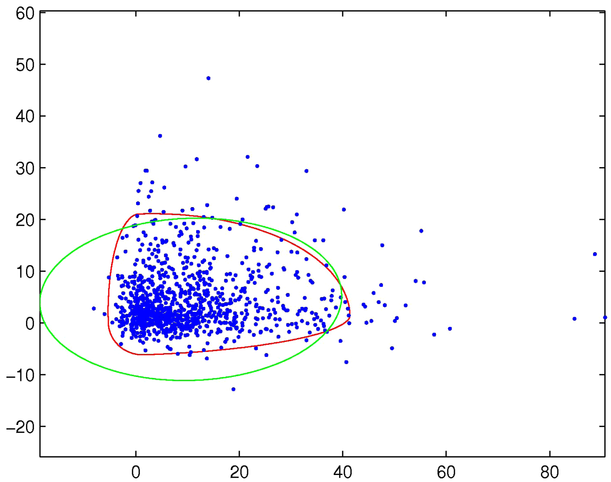

Fig. 1. Skew error ellipse versus standard error ellipse: The skew error ellipse is indicated in red while the standard error ellipse is in green (color online).

Proof. The probability of being within the skew error ellipse with axes $c\sigma_{X_-}$, $c\sigma_{X_+}$, $c\sigma_{Y_-}$ and $c\sigma_{Y_+}$ equals

$$\begin{aligned}&\int_0^{c\sigma_{Y_+}}\int_0^{c\sigma_{X_+}\sqrt{1-(y/c\sigma_{Y_+})^2}} g_X(x)g_Y(y)dxdy\\ &\quad+\int_0^{c\sigma_{Y_+}}\int_{c\sigma_{X_-}\sqrt{1-(y/c\sigma_{Y_+})^2}}^0 g_X(x)g_Y(y)dxdy\\ &\quad+\int_{c\sigma_{Y_-}}^0\int_{c\sigma_{X_-}\sqrt{1-(y/c\sigma_{Y_-})^2}}^0 g_X(x)g_Y(y)dxdy\\ &\quad+\int_{c\sigma_{Y_-}}^0\int_0^{c\sigma_{X_+}\sqrt{1-(y/c\sigma_{Y_-})^2}} g_X(x)g_Y(y)dxdy\\ &\quad=\left(\frac{2}{k_X^2+1}\right)\left(\frac{2}{k_Y^2+1}\right)\left(\frac{1}{4}\alpha+\frac{k_X^2}{4}\alpha+\frac{k_Y^2}{4}\alpha+\frac{k_X^2k_Y^2}{4}\alpha\right)=\alpha.\end{aligned}$$

□

For the remainder of this discourse, one may assume that the data has mode zero in each coordinate with sample standard deviation in each coordinate being σ_1 and σ_2. Consider an error ellipse with axes having values σ_1 and σ_2 and center at the mean of the data. Then the area of the error ellipse is $\pi\sigma_1\sigma_2$. Let k_X, σ_X and k_Y, σ_Y be the maximum likelihood estimators of the x- and y-coordinates, respectively, for whichever distribution is used to model skew in each coordinate. The area of the associated skew ellipse with center at the mode (assumed to be zero) will be

$$\begin{aligned}&\frac{\pi}{4}\sigma_{X_+}\sigma_{Y_+}+\frac{\pi}{4}\sigma_{X_-}\sigma_{Y_+}+\frac{\pi}{4}\sigma_{X_-}\sigma_{Y_-}+\frac{\pi}{4}\sigma_{X_+}\sigma_{Y_-}\\ &\quad=\frac{\pi}{4}k_Xk_Y\sigma_X\sigma_Y+\frac{\pi}{4}(k_X/k_Y)\sigma_X\sigma_Y+\frac{\pi}{4}(1/k_Xk_Y)\sigma_X\sigma_Y+\frac{\pi}{4}(k_Y/k_X)\sigma_X\sigma_Y\end{aligned}$$

$$= \frac{\pi}{4}\left(\frac{k_X^2+1}{k_X}\right)\left(\frac{k_Y^2+1}{k_Y}\right)\sigma_X\sigma_Y.$$

Our goal is to compare the areas of the error ellipses, namely

$$\frac{\pi}{4}\left(\frac{k_X^2+1}{k_X}\right)\left(\frac{k_Y^2+1}{k_Y}\right)\sigma_X\sigma_Y \quad \text{and} \quad \pi\sigma_1\sigma_2.$$

We now provide the statement of the culminating result. The proof has been relegated to Appendix B.

Corollary 1. *If the two-dimensional data is symmetric around zero, then*

(1) The area of the skew normal error ellipse equals the area of the standard error ellipse.
(2) The skew Laplace error ellipse is less than or equal to twice the area of the standard error ellipse.

Due to the nature in which $\hat{k}$ and $\hat{a}$ are found, it is difficult to prove a Corollary 1-type result for skewed generalized Cauchy distributions, however empirical evidence suggests that the area is quite comparable to the area of the standard error ellipse. Note also that Corollary 1 easily generalizes to hyper-ellipses of arbitrary dimension.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

5. Towards Classification

5.1. *Skew distance*

One classic binary classifier uses Mahalanobis distances to make a decision as to which class a test point belongs. Distance from a point to the mean of a collection of points is defined as follows:

Definition: Let $\mathbf{x} \in \mathbb{R}$ and $X \subseteq \mathbb{R}^d$ be a collection of points with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2\ \cdots\ \mu_d]$. Let A be the matrix whose rows are the points in X. Let E be a square matrix with columns forming the PCA basis for A. Let $B = EA$ and set $\mathbf{y} = [y_1\ y_2\ \cdots\ y_d] = E\mathbf{x}^T$. Consider the standard deviation in each of the PCA directions, σ_i $(i = 1, 2, \ldots, d)$. The squared distance from $\mathbf{x}$ to X is defined to be

$$d(\mathbf{x}, X) = \sum_{i=1}^{d} \frac{(y_i - \mu_i)^2}{\sigma_i^2}.$$

We will now generalize this distance to the case where we take into consideration the skew of the data.

Definition: Let $\mathbf{x} \in \mathbb{R}$ and $X \subseteq \mathbb{R}^d$ be a collection of points which we may assume has mean $\mathbf{0}$. Let A be the matrix A whose rows are the points in X. Let E be a square matrix with columns forming the PCA basis for A. Let $B = EA$ and set $\mathbf{y} = [y_1\ y_2\ \cdots\ y_d] = E\mathbf{x}^T$. Let the rows of B have mode $\mathbf{m} = [m_1\ m_2\ \cdots\ m_d]$ and, as defined above, consider the negative and positive standard deviations $(\sigma_i)_-$, $(\sigma_i)_+$ $(i = 1, 2, \ldots, d)$ corresponding to the associated skew distribution and its maximum likelihood estimators. The squared distance from $\mathbf{x}$ to the mode of X is defined to be

$$d(\mathbf{x}, X) = \sum_{y_i < m_i} \frac{(y_i - m_i)^2}{(\sigma_i^2)_-} + \sum_{y_i \geq m_i} \frac{(y_i - m_i)^2}{(\sigma_i^2)_+}. \tag{1}$$

Note that when the data has no skew this new skew distances defined reduces to Mahalanobis distance. We wish to point out that in order to compute the mode for continuous data we opted to use kernel density estimation and chose the maximizer for the resulting density function.

5.2. *A classifier based on skew distance*

Given two classes $C_1, C_2 \subseteq \mathbb{R}^d$ of training data, we now give a top-down outline of the algorithm for training the skew binary classifier:

(1) Perform the Skew Distance Transformation from $C_1 \cup C_2$ into $\mathbb{R}^2$. In order to perform this transformation, several things must be done for each class C_i $(i = 1, 2)$:

   (a) Find the mean of class C_i and shift so that its mean is zero.
   (b) Compute the principal components of C_i and perform the corresponding change of basis. Call the resulting set P_i.
   (c) Find the mode of P_i and shift so that its mode is zero. Call the resulting set Q_i.
   (d) In each coordinate of Q_i, fit the data with a skew distribution using maximum likelihood estimation according to this criterion:

      (i) If skew ≤ 1, then fit the data with a skew normal distribution.
      (ii) If $1 <$ skew ≤ 2, then fit the data with a skew Laplace distribution.
      (iii) If skew > 2, then fit the data with a skew generalized Cauchy distribution.

   (e) Use this pdf to determine the one-sided standard deviations.
   (f) For each point $\mathbf{x} \in P_i$ associate a point in $(d(\mathbf{x}, P_1), d(\mathbf{x}, P_2)) \in \mathbb{R}^2$ according to the definition provided in Eq. (1). Set $T_i = \{(d(\mathbf{x}, P_1), d(\mathbf{x}, P_2)) \mid \mathbf{x} \in P_i\}$.

(2) Compute the Minimal Total Error Fisher Linear Discriminant on the transformed classes T_1 and T_2:

   (a) Compute the classic Fisher Linear Discriminant. Let the slope be m.
   (b) Find the line with slope m which minimizes the total error in classifying both classes.

In Fig. 2, we applied this algorithm to the Ionosphere dataset found in the UCI Machine Learning Repository.

5.3. *Experimental results*

The classifier was applied to seven data sets from the UCI Machine Learning Repository.[3] We chose numerical data sets with no missing values consisting of two classes. Attention was made to selecting data sets which had marked skew (Table 1).

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

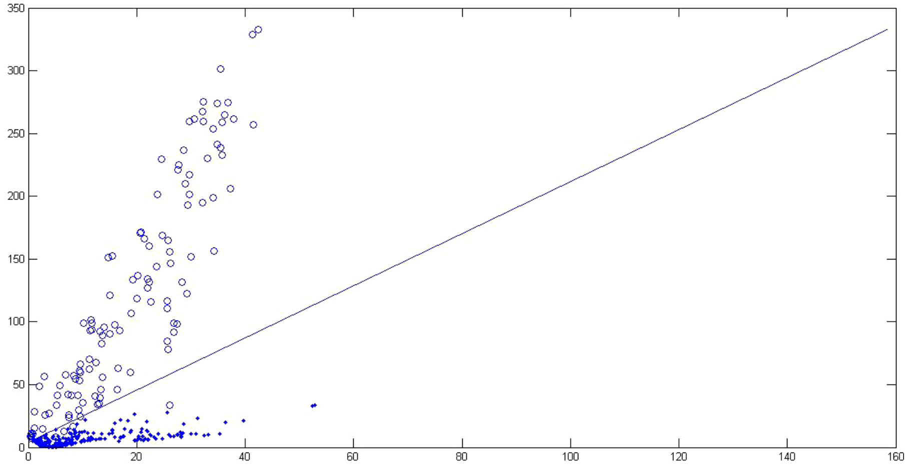

Fig. 2. Skew transform on ion data with MTE Fisher line.

Consider the case of the *liver* data set. After shifting the mean to zero, rotating the data to its principle components, and shifting the mode to zero we computed skew in the components of class one and class two. Table 2 lists the results.

One sees that the sixth component in both classes exhibits a high amount of skew and this fact contributes to the success of our classifier. In Fig. 3, we created a histogram of the sixth component of each class.

We performed 10-fold validation 10 times (thus, 100 instances) and computed the mean and standard deviation of the error rate.

First, we wanted to verify that using skew distance can indeed improve classification versus other distance metrics. Therefore, we implemented our classifier

Table 1. UCI machine learning repository data sets implemented.

UCI data	Instances	Attributes	Classes
breast	569	30	2
ionosphere	351	32	2
diabetes	768	8	2
heart	120	44	2
liver	341	6	2
appendicitus	106	7	2
sonar	208	60	2

Table 2. Computed skew for each of the six components of each class of the liver data set.

Class	c_1 Skew	c_2 Skew	c_3 Skew	c_4 Skew	c_5 Skew	c_6 Skew
1	0.4	0	0.3	1.1	0.6	-2.6
2	0.6	-0.5	-2.1	0.6	1.4	-2.5

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

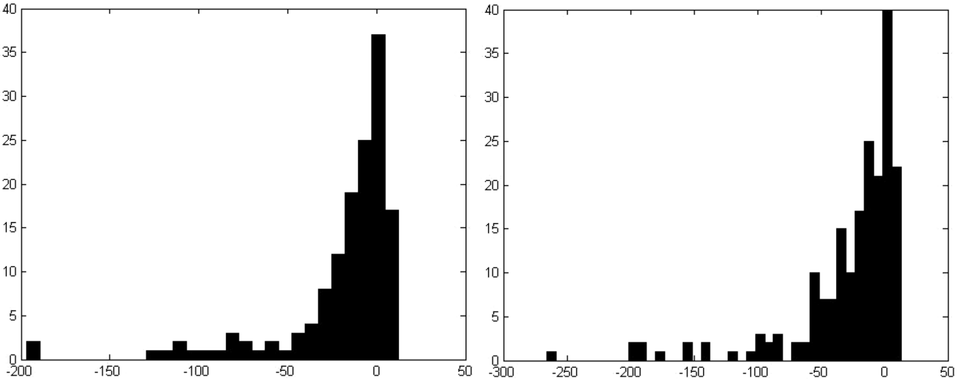

Fig. 3. Histograms of the sixth component of each class of the liver data set.

algorithm described in the previous section but varied the distance function over different distance metrics. For nearly every data set tested skew distance improved (or in the case of Mahalanobis distance, had little or no negative effect on) the classifier. There was one case in which normal Euclidean distance outperformed the rest. Table 3 summarizes our results.

We now list our results using our skew classifier and show that they are comparable to results in other papers citing these data sets using other popular methods.

In Ref. 8, using 10-fold cross-validation, the authors compare their proposed classifier, support feature machines (SFM) to several popular classification methods which include support vector machines (SVM). In Ref. 9, again using 10-fold cross-validation 10 times, the authors compare their proposed classifier (oRF) to several popular classification methods which include random forest (RF) algorithms. Using the same experimental method, Table 4 shows the comparison of our results to some of the results listed in Refs. 8 and 9. Our results are comparable and in two cases better than the other methods.

In Ref. 6, using 10 iterations of 80% training and 20% testing, the authors compare their proposed classifier (CMTNN) to several popular neural network classification methods. Using the same experimental method, Table 5 shows the comparison of our results to the results listed in Ref. 6. Our results are comparable and in one instance better than the other methods.

Table 3. % error rate results: mean ± standard deviation on UCI data using skew versus other distance metrics.

UCI data	Skew	Mahalanobis	Euclidean	Manhattan	Chebyshev
breast	**11.2** ± 3.9	11.8 ± 5.0	39.7 ± 7.5	35.5 ± 7.8	34.9 ± 6.9
ionosphere	**4.4** ± 3.4	9.8 ± 4.9	26.1 ± 6.7	22.8 ± 7.8	21.9 ± 6.4
diabetes	29.4 ± 4.6	**27.9** ± 4.9	38.6 ± 7.2	31.3 ± 5.4	37.3 ± 5.3
heart	**8.8** ± 4.7	42.0 ± 9.8	28.2 ± 8.2	25.1 ± 7.1	44.4 ± 7.4
liver	35.9 ± 8.4	**35.4** ± 7.9	39.8 ± 7.6	44.0 ± 8.3	42.5 ± 7.9
appendicitus	17.1 ± 10.7	24.2 ± 13.7	**11.2** ± 11.4	13.8 ± 9.1	31.3 ± 13.8
sonar	**19.8** ± 8.1	30.5 ± 9.7	32.8 ± 12.3	24.1 ± 9.9	43.4 ± 9.7

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

Table 4. % error rate results: A comparison of our proposed method versus methods found in Refs. 8 and 9 on UCI data.

UCI data	Skew	SVML[8]	SVMG[8]	SFM[8]	kNN[9]	CART[9]	adaboost[9]	oRF-rnd[9]	oRF-lda[9]	oRF-ridge[9]
ionosphere	**4.4**	10.5	5.4	5.4	14	10.5	6.1	5.6	5.4	5.6
diabetes	29.4	23.1	23.8	**22.4**	31.7	34.2	26.6	27.7	26.1	26.2
heart	**8.8**	17.5	17.2	18.8	35.5	21.3	18.3	20	18	17.9
liver	35.9	—	—	—	32	34.1	26.8	27.8	**25.8**	26.2
appendicitus	17.1	**12.4**	13.3	13.2	—	—	—	—	—	—
sonar	19.8	24.5	13.4	**12.0**	18.2	27.2	13.1	14.4	18	18.2

Table 5. % accuracy results (mean ± standard deviation): A comparison of our proposed method versus methods found in Ref. 6 on UCI data.

UCI data	Skew	BPNN[6]	GRNN[6]	RBFNN[6]	PNN[6]	CMTNN[6]
ionosphere	**94.4** ± 2.7	90.3 ± 4.2	93.1 ± 1.8	90.1 ± 4.0	85.6 ± 4.0	93.4 ± 2.9
diabetes	72.1 ± 3.4	76.2 ± 4.4	75.3 ± 3.9	**76.6** ± 2.5	75.3 ± 3.9	76.5 ± 3.4
liver	64.2 ± 5.3	70.0 ± 6.3	64.1 ± 6.7	67.5 ± 4.5	64.1 ± 6.7	**70.7** ± 7.3

6. Conclusion

We have shown in this paper that one might benefit from incorporating skew into a classifier. Table 3 suggests that this might be the case, for when we added skew to the mean and standard deviation of data, the resulting classifier demonstrated in some cases improvement over other distance metrics. Tables 4 and 5 illustrate that a classifier based on skew distance alone can have results comparable and sometimes better than known classifiers. Our classifier works well on data sets which exhibit ample enough skew. In the course of the experimentation we found that it does not perform as well on categorical data in which the categories have been replaced by whole numbers. We believe that a classifier which makes use of a distance metric could benefit from considering the use of this skew distance metric. Incorporating other techniques into our classifier could only make it better and we shall do this in future papers. For example, perhaps kurtosis could be incorporated into the distance metric or a distance metric that takes into account the clusters of each class (distance to a class could be defined as distance to the closest cluster in the class). However, our intention for writing this paper was to simply propose the use of skew in classification. The results show that this just might be worth considering.

Acknowledgments

The author would like to thank Andrew Ippoliti, Eric Mason and Alex Stein for their invaluable help in preparing the data and writing some of the code for evaluating the merit of this new proposed classifier.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

Appendix A. Proof of Some Relevant Facts

In Ref. 4, it is shown that the nth-order moment for the skewed distribution is

$$M_n\left(\frac{k^{n+1}+\frac{(-1)^{n+1}}{k^{n+1}}}{k+\frac{1}{k}}\right), \quad \text{where } M_n = 2\int_0^\infty x^n f(x)dx.$$

It is pointed out in Ref. 4, and it is easy to verify, that in general the skewness of $g(x)$ is bounded by the skew of one tail of the symmetric distribution $f(x)$. Using this information, we see that the statistics for the skewed generalized cauchy distribution are as follows:

$$\begin{aligned}
\mu &= \frac{2k^2(k^2-1)}{(k^2+1)(a-2)}, \quad \text{for } a>2, \quad \text{mode}=0,\\
\sigma^2 &= \frac{4k^2[k^6+2(a-3)k^4-2(a-3)k^2+(a-2)]}{(a-3)(a-2)^2(k^2+1)^2}, \quad \text{for } a>3,\\
\text{skew} &= \frac{\begin{array}{c}-\sqrt{a-3}(k^2-1)[(a-6)(a+2)k^8+2(a-6)(5a-14)k^6\\ -\,2(7a^2-40a+60)k^4+6(a-6)(a-2)k^2-3(a-2)^2]\end{array}}{4(a-4)k[k^6+2(a-3)k^4-(a-3)k^2+(a-2)]^{3/2}}, \quad \text{for } a>4.
\end{aligned}$$

Note that the skew can be made as large as we like by letting a tend to 4 from the right in the formula above.

Theorem A.1. *Consider the data $x_1, x_2, \ldots, x_N$. Then*

(1) *When k is known, the maximum likelihood estimator*

$$\hat{a} = \frac{N}{\sum_{x_i<0}\ln(1+k|x_i|)+\sum_{x_i\geq 0}\ln(1+\frac{|x_i|}{k})}+1.$$

(2) *When a is known, the maximum likelihood estimator $\widehat{k}$ is the unique point of intersection of the following two curves:*

$$y_1(k) = \sum_{x_i<0}\frac{kx_i}{1-kx_i}+\sum_{x_i\geq 0}\frac{\frac{x_i}{k}}{1+\frac{x_i}{k}} \quad \text{and} \quad y_2(k)=\frac{N(k^2-1)}{a(k^2+1)}.$$

(3) *When both a and k are unknown, the maximum likelihood estimator $\widehat{k}$ is the unique point of intersection of the curves*

$$\begin{aligned}
y_3(k) &= \frac{\frac{N(k^2-1)}{k^2+1}}{\sum_{x_i<0}\frac{kx_i}{1-kx_i}+\sum_{x_i\geq 0}\frac{(x_i/k)}{1+(x_i/k)}} \quad \text{and}\\
y_4(k) &= \frac{N}{\sum_{x_i<0}\ln(1-kx_i)+\sum_{x_i\geq 0}\ln(1+\frac{x_i}{k})}+1.
\end{aligned}$$

The value of $\hat{a}$ is then obtained as in part (1) of this theorem.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

Proof. The log-likelihood function in this case has the form

$$\begin{aligned}\ell(a,k) &= N[\ln(a-1)+\ln k-\ln(1+k^2)]\\ &\quad - a\left[\sum_{x_i<0}\ln(1-kx_i)+\sum_{x_i\geq 0}\ln\left(1+\frac{x_i}{k}\right)\right].\end{aligned}$$

Solving the equation $\ell_a = 0$ for the variable a, we have

$$\hat{a} = \frac{N}{\sum_{x_i<0}\ln(1-kx_i)+\sum_{x_i\geq 0}\ln(1+\frac{x_i}{k})}+1.$$

Thus, when k is known the above formula gives us the estimator for a.

Let us assume now that a is known. Computing the other partial derivative, we have

$$\ell_k = N\left(\frac{1}{k}-\frac{2k}{1+k^2}\right) - a\left[\sum_{x_i<0}\frac{-x_i}{(1-kx_i)}+\sum_{x_i\geq 0}\frac{-x_i/k^2}{(1+x_i/k)}\right].$$

We solve the equation $\ell_k = 0$ for the variable k by first multiplying through by k/a to get the equation

$$\sum_{x_i<0}\frac{kx_i}{1-kx_i}+\sum_{x_i\geq 0}\frac{\frac{x_i}{k}}{1+\frac{x_i}{k}} = \frac{N(k^2-1)}{a(k^2+1)}.$$

Thus, we see that k is a point of intersection of the following two curves:

$$y_1(k) = \sum_{x_i<0}\frac{kx_i}{1-kx_i}+\sum_{x_i\geq 0}\frac{\frac{x_i}{k}}{1+\frac{x_i}{k}} \quad \text{and} \quad y_2(k) = \frac{N(k^2-1)}{a(k^2+1)}.$$

Now we explain why there is unique solution. One can check that $y_1(k)$ is strictly decreasing, while $y_2(k)$ is strictly increasing. Indeed,

$$\begin{aligned}y_1'(k) &= \sum_{x_i<0}\frac{x_i}{(1-kx_i)^2}+\sum_{x_i\geq 0}\frac{-x_i}{(1+\frac{x_i}{k})^2} < 0, \quad \text{while}\\ y_2'(k) &= \frac{4Nk}{a(k^2+1)^2} > 0.\end{aligned}$$

These two curves must intersect in a unique point, since $y_1(k)$ is initially positive and becomes negative as k tends to infinity, while $y_2(k)$ is initially negative and becomes positive as k tends to infinity.

Finally, we consider the case when both a and k are unknown. Take the equation generated from $\ell_k = 0$ and now solve for a to get

$$a = \frac{\frac{N(k^2-1)}{k^2+1}}{\sum_{x_i<0}\frac{kx_i}{1-kx_i}+\sum_{x_i\geq 0}\frac{(x_i/k)}{1+(x_i/k)}}.$$

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

Let us define the follow functions:

$$y_3(k) = \frac{\frac{N(k^2-1)}{k^2+1}}{\sum_{x_i<0} \frac{kx_i}{1-kx_i} + \sum_{x_i\geq 0} \frac{(x_i/k)}{1+(x_i/k)}} \quad \text{and}$$

$$y_4(k) = \frac{N}{\sum_{x_i<0} \ln(1-kx_i) + \sum_{x_i\geq 0} \ln(1+\frac{x_i}{k})} + 1.$$

Equating the two expressions we have for a in terms of k we show that $y_3(k) = y_4(k)$ has a unique solution.

Indeed, $y_3(k)$ is the quotient of $y_2(k)$ (with $a = 1$) over $y_1(k)$, and so is strictly increasing for $k > 1$. Now $y_4(k)$ is first increasing and then decreasing, since

$$y_4'(k) = \frac{Ny_1(k)}{k\big(\sum_{x_i<0} \ln(1-kx_i) + \sum_{x_i\geq 0} \ln(1+\frac{x_i}{k})\big)^2},$$

and so the sign of $y_4'(k)$ is determined by the sign of $y_1(k)$ which we have already determined above.

Furthermore, $y_4(k)$ tends towards 1 as k tends towards 0 or infinity. □

Appendix B. Proof of Corollary 1

In order to prove Corollary 1, we will first need a few technical results. The first is a straightforward computation, so we omit the proof.

Lemma B.1. *Consider one-dimensional data* $x_1, x_2, \ldots, x_N$.

(1) *The maximum likelihood estimators for the skewed normal distribution are*

$$\hat{\sigma} = \sqrt{\frac{ak^2 + b(1/k^2)}{N}} \quad \textit{and} \quad \hat{k} = (b/a)^{1/6}, \quad \textit{where}$$

$$a = \sum_{x_i<0} x_i^2 \quad \textit{and} \quad b = \sum_{x_i\geq 0} x_i^2.$$

(2) *The maximum likelihood estimators for the skewed Laplace distribution can be shown to be*

$$\hat{\lambda} = \frac{ak + b(1/k)}{N} \quad \textit{and} \quad \hat{k} = (b/a)^{1/4}, \quad \textit{where}$$

$$a = \sum\nolimits_{x_i<0} |x_i| \quad \textit{and} \quad b = \sum\nolimits_{x_i\geq 0} |x_i|.$$

Lemma B.2. *Consider one-dimensional data* $x_1, x_2, \ldots, x_N$ *with mean equal to zero.*

(1) *If* k *and* σ *are the maximum likelihood estimators of the skew normal distribution for this data, then*

$$\left(\frac{k^2+1}{k}\right)\sigma = \sqrt{\frac{(\sqrt[3]{a} + \sqrt[3]{b})^3}{N}},$$

where $a = \sum_{x_i<0} x_i^2$ *and* $b = \sum_{x_i\geq 0} x_i^2$.

(2) *If k and λ are the maximum likelihood estimators of the skew Laplace distribution for this data, then*

$$\left(\frac{k^2+1}{k}\right)\sigma = \frac{\sqrt{2}(\sqrt{a}+\sqrt{b})^2}{N},$$

where $a = \sum_{x_i<0}|x_i|$ and $b = \sum_{x_i\geq 0}|x_i|$.

Proof. For the first item, note that

$$\begin{aligned}\left(\frac{k^2+1}{k}\right)\sigma &= \left(\frac{k^2+1}{k}\right)\sqrt{\frac{ak^2+b(1/k^2)}{N}} = \frac{1}{\sqrt{N}}\frac{(k^2+1)\sqrt{ak^4+b}}{k^2}\\ &= \frac{1}{\sqrt{N}}\left(\left(\frac{b}{a}\right)^{1/3}+1\right)\left(\frac{a}{b}\right)^{1/3}\sqrt{a\left(\frac{b}{a}\right)^{2/3}+b}\\ &= \frac{1}{\sqrt{N}}(a^{1/3}+b^{1/3})\sqrt{a^{1/3}+b^{1/3}} = \sqrt{\frac{(a^{1/3}+b^{1/3})^3}{N}}.\end{aligned}$$

For the second item, note that

$$\begin{aligned}\left(\frac{k^2+1}{k}\right)\sigma &= \left(\frac{k^2+1}{k}\right)\sqrt{2}\left(\frac{ak+b(1/k)}{N}\right) = \frac{\sqrt{2}[ak^4+(a+b)k^2+b]}{k^2N}\\ &= \frac{\sqrt{2}\left[a\left(\frac{b}{a}\right)+(a+b)\sqrt{\frac{b}{a}}+b\right]}{\sqrt{\frac{b}{a}}N} = \frac{\sqrt{2}\left[a\sqrt{\frac{b}{a}}+(a+b)+b\sqrt{\frac{a}{b}}\right]}{N}\\ &= \frac{\sqrt{2}(a+2\sqrt{ab}+b)}{N} = \frac{\sqrt{2}(\sqrt{a}+\sqrt{b})^2}{N}.\end{aligned}$$

□

Lemma B.3. *For any non-negative real numbers a and b, we have*

$$\frac{(\sqrt[n]{a}+\sqrt[n]{b})^n}{2^{n-1}} \leq a+b \quad \text{for } n = 2,3,4,\ldots.$$

Equality exists iff $a = b$.

Proof. Consider the function

$$f(a,b) = \frac{(\sqrt[n]{a}+\sqrt[n]{b})^n}{2^{n-1}} - a - b.$$

We show that f has all its global maxima when $a = b$ (in which case $f(a,b) = f(a,a) = 0$). We may assume that $a, b \neq 0$. Indeed, for instance, if $a \neq 0$ but $b = 0$, we have $f(a,b) = -(\frac{2^{n-1}-1}{2^{n-1}})a < 0$. Now, one can compute that

$$f_a = \frac{1}{2^{n-1}}(1+(b/a)^{1/n})^{n-1} - 1, \quad f_b = \frac{1}{2^{n-1}}(1+(a/b)^{1/n})^{n-1} - 1.$$

From this it follows that the critical points of f consists of all the points on the line $a = b$. Indeed, if we solve $f_a = 0, f_b = 0$ by setting $u = (b/a)^{n-1}$, we arrive at $(1+u)^{n-1} = 2^{n-1}$ and $(1+(1/u))^{n-1} = 2^{n-1}$ and so $u = 1$. Thus, $a = b$. One can

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

now compute that

$$\begin{aligned} f_{aa} &= -\frac{n-1}{n2^{n-1}a}(b/a)^{1/n}(1+(b/a)^{1/n})^{n-2} \quad \text{and} \\ f_{aa}(a,a) &= -\frac{n-1}{2na} < 0, \quad \text{while } f_{aa}(a,a)f_{bb}(a,a) - f_{ab}(a,a)^2 = 0. \end{aligned}$$

Hence, f is concave on its restricted domain, the first quadrant, which makes the critical points (a,a), for $a > 0$, global maxima. In other words, for all $a, b \geq 0$ we have $f(a,b) \leq f(a,a) = 0$, and this yields the required inequality.

Furthermore, for $a \neq b$, we have $f(a,b) < f(a,a) = 0$, and for $a = b$ we have $f(a,a) = 0$. These two statements prove that

$$\frac{(\sqrt[n]{a}+\sqrt[n]{b})^n}{2^{n-1}} = a + b \quad \text{iff } a = b.$$

□

Theorem B.1. *The following statements hold for a any collection of two-dimensional data* $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$:

(1) *The area of the skew normal ellipse is less than or equal to*

$$\pi\sqrt{\frac{\sum_{i=1}^N x_i^2}{N}}\quad\sqrt{\frac{\sum_{i=1}^N y_i^2}{N}},$$

with equality iff

$$\sum_{x_i<0} x_i^2 = \sum_{x_i\geq 0} x_i^2.$$

(2) *The area of the skew Laplace ellipse is less than or equal to*

$$2\pi\left(\frac{\sum_{i=1}^N |x_i|}{N}\right)\left(\frac{\sum_{i=1}^N |y_i|}{N}\right),$$

with equality iff

$$\sum_{x_i<0} |x_i| = \sum_{x_i\geq 0} |x_i|.$$

Proof. Recall that the area of any skew error ellipse is

$$\frac{\pi}{4}\left(\frac{k_X^2+1}{k_X}\right)\left(\frac{k_Y^2+1}{k_Y}\right)\sigma_X\sigma_Y.$$

To prove this theorem, it suffices to show that given a set of data $x_1, x_2, \ldots, x_N$ with maximum likelihood estimators k_X and σ_X for the given skew distribution, we have the appropriate bound on

$$\left(\frac{k_X^2+1}{k_X}\right)\sigma_X.$$

For the first item, set $a_X = \sum_{x_i<0} x_i^2$ and $b_X = \sum_{x_i \geq 0} x_i^2$. By the previous two lemmas,

$$\left(\frac{k_X^2+1}{k_X}\right)\sigma_X = \sqrt{\frac{(\sqrt[3]{a_X}+\sqrt[3]{b_X})^3}{N}} \leq 2\sqrt{\frac{a_X+b_X}{N}} = 2\sqrt{\frac{\sum_{i=1}^N x_i^2}{N}},$$

with equality iff $a_X = b_X$. For the second item, set $a_X = \sum_{x_i<0}|x_i|$ and $b_X = \sum_{x_i \geq 0}|x_i|$. By the previous two lemmas,

$$\left(\frac{k_X^2+1}{k_X}\right)\sigma_X = \frac{\sqrt{2}(\sqrt{a_X}+\sqrt{b_X})^2}{N} \leq \frac{2\sqrt{2}(a_X+b_X)}{N} = \frac{2\sqrt{2}\sum_{i=1}^N |x_i|}{N},$$

with equality iff $a_X = b_X$. □

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com by Dr. Mark DeBonis on 04/08/15. For personal use only.

Now we prove Corollary 1:

First note that if the data is symmetric around zero, then certainly the mean is zero. For the first item we would have $\sum_{x_i<0} x_i^2 = \sum_{x_i \geq 0} x_i^2$, and so the area of the skew normal ellipse equals

$$\pi\sqrt{\frac{\sum_{i=1}^N x_i^2}{N}}\sqrt{\frac{\sum_{i=1}^N y_i^2}{N}},$$

which is also the area of the standard error ellipse. For the second item we would have $\sum_{x_i<0}|x_i| = \sum_{x_i \geq 0}|x_i|$, and so the area of the skew Laplace error ellipse equals

$$2\pi\left(\frac{\sum_{i=1}^N |x_i|}{N}\right)\left(\frac{\sum_{i=1}^N |y_i|}{N}\right).$$

Since $E[|X|] \leq \sqrt{E[X^2]}$, this is less than or equal to

$$2\pi\sqrt{\frac{\sum_{i=1}^N x_i^2}{N}}\sqrt{\frac{\sum_{i=1}^N y_i^2}{N}},$$

which is twice the area of the standard error ellipse.

References

1. R. B. Arellano-Valle, M. D. Branco and M. G. Genton, A unified view on skewed distributions arising from selections, *Can. J. Stat.* **34**(4) (2006) 581–601.
2. A. Azzalini, The skew-normal distribution and related multivariate families, *Scand. J. Stat.* **32**(2) (2005) 159–188.
3. C. Blake, E. Keogh and C. Merz, UCI repository of machine learning databases. Department of Information and Computer Science, UC, Irvine (1998), Available at http://wwwicsuciedu/mlearn/MLRepositoryhtml.
4. C. Fernandez and M. F. Steel, On bayesian modeling of fat tails and skewness, *J. Am. Stat. Assoc.* **93** (1998) 359–371.
5. M. G. Genton (ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality* (CRC Press, 2004).

6. P. Jeatrakul and K. W. Wong, Comparing the performance of different neural networks for binary classification problems, *SNLP '09. Eighth Int. Symp. Natural Language Processing, 20–22 October 2009*, pp. 111–115.
7. Y. Li, L. Q. Xu, J. Morphett and R. Jacobs, An integrated algorithm of incremental and robust PCA, in *Proc. IEEE Int. Conf. Image Processing (ICIP2003)*, Barcelona, Spain, September 2003, 1-245-8, Vol. 1.
8. T. Maszczyk and D. Wlodzislaw, Support feature machines: Support vectors are not enough, *WCCI 2010 IEEE World Congress on Computational Intelligence, CCIB*, Barcelona, Spain, 18–23, July 2010, pp. 3852–3859.
9. B. Menze, M. Kelm, D. Splitthoff, U. Koethe and F. Hamprecht, On oblique random Forests, *Machine Learning and Knowledge Discovery in Databases* (Springer, 2011), pp. 435–469.
10. P. R. Rider, Generalized cauchy distributions, *Ann. Inst. Stat. Math.* **9** (1957) 215–223.
11. N. Sebe, M. S. Lew, I. Cohen, A. Garg and T. S. Huang, Emotion recognition using a Cauchy Naive Bayes classifier, in *Proc. ICPR* (Sensors, 2002), pp. 3852–3859.
12. S. Xu, Z. Zhou, H. Lu, X. Luo and Y. Lan, Improved algorithms for the classification of rough rice using a bionic electronic nose based on PCA and the Wilks distribution, *Sensors* **14**(3) (2014) 5486–5501.

Int. J. Patt. Recogn. Artif. Intell. Downloaded from www.worldscientific.com
by Dr. Mark DeBonis on 04/08/15. For personal use only.

Mark DeBonis received his PhD in Mathematics in 1991 from University of California, Irvine, USA. He spent some time working for the US Department of Energy and Department of Defense as an applied mathematician on applications of machine learning. Currently, he is an Assistant Professor at Manhattan College in New York City.

His research interests include machine learning, statistics, computational algebra and cryptology.