Comparación de las metodologías de análisis discriminante robusto y redes neuronales

*Héctor René Álvarez L.**
*Gerardo Avendadño P.***

Fecha de recepción: 11 de agosto de 2014
Fecha de aprobación: 2 de octubre de 2014
Pp. 35-64

Resumen

En muchas aplicaciones en ingeniería y otras áreas de conocimiento es frecuente que los investigadores tengan que enfrentarse con la necesidad de identificar las características que permitan diferenciar a dos o más grupos de individuos u objetos, y casi siempre esta situación está motivada por la necesidad de poder clasificar nuevos casos como pertenecientes a uno u otro grupo. En el artículo se analiza el problema de la discriminación de una población cuando tienen observaciones atípicas (*outliers*), aplicando la metodología de redes neuronales (clasificación robusta utilizando redes neuronales), y su objetivo es compararla con la metodología de discriminación robusta, analizando sus ventajas.

Palabras clave

Redes neuronales, *Outliers*, discriminación robusta, clasificación de observaciones.

* Ph.D. en Aplicaciones Técnicas e Informáticas de la Estadística, Universidad Politécnica de Cataluña.

** Postdoctorado en Manufacturing Engenniering, Penn State University. Ph.D. en Métodos Estadísticos Avanzados, Universidad Politécnica de Valencia.

Comparison of Different Methodologies in the Robust Discriminating Analysis and Neuronal Networks

Abstract

In lots of applications of engineering and other related areas, it is frequently seen that researchers have to face the need of identifying specific features which make them differentiate two or more groups of individuals or objects, and very often this situation is caused by the need to classify new cases as belonging to one group or another. This research paper aims at studying the discrimination of a particular population which shows atypical observation results (Outliers) with the application of neuronal network methodology, a robust classification using neuronal network, and whose main objective is to compare it with a robust discrimination methodology, analyzing its advantages.

Key words

Neuronal network, Outliers, robust discrimination, classification of observation results.

Comparaison des méthodologies d'analyse discriminante robuste et réseaux neuronaux

Résumé

Il est fréquent de constater que dans nombre d'applications d'ingénierie mais également dans d'autres domaines de la connaissance, les chercheurs doivent faire face à la nécessité d'identifier les caractéristiques qui différencient deux ou plusieurs groupes de personnes ou d'objets. Cette situation est presque toujours motivée par la nécessité de pouvoir classer les nouveaux cas comme appartenant à l'un des deux groupes. L'article analyse le problème de la discrimination d'une population lorsqu'il existe des observations atypiques (valeurs aberrantes) et applique la

méthodologie de réseaux neuronaux (classification robuste utilisant les réseaux de neurones), L'objectif est alors de comparer cette population avec la méthodologie de discrimination robuste et d'en analyser ses avantages.

Mots-clés

réseaux neuronaux, valeurs aberrantes, discrimination robuste, classification des observations

Comparação de metodologias para análise discriminante robusta e redes neurais

Resumo

Em muitas aplicações de engenharia e outras áreas do conhecimento prevalece que os investigadores enfrentem a necessidade de identificar características que permitam diferenciar dois ou mais grupos de indivíduos ou objetos, e muitas vezes esta situação é motivada pela necessidade de poder classificar novos casos como pertencentes a um grupo ou outro. No artigo o problema da discriminação de uma população quando analisados com observações atípicas (outliers) aplicando a metodologia de redes neurais (classificação robusta utilizando redes neurais), e seu objetivo é compara-la a metodologia de discriminação robusta, analisando as suas vantagens.

Palavras-chave

Redes neurais, Outliers, discriminação robusta, classificação de observações.

1. Introducción

En muchas aplicaciones en ingeniería y otras áreas de conocimiento es frecuente que los investigadores tengan que enfrentarse con la necesidad de identificar las características que permitan diferenciar a dos o más grupos de individuos u objetos. Casi siempre, esta situación está motivada por la necesidad de poder clasificar nuevos casos como pertenecientes a uno u otros grupos.

Por ejemplo, entre otros, en el análisis de riesgos financieros, se plantea determinar si un cliente está bien clasificado como riesgoso para aprobar un préstamo; en el caso de selección de personal, la pregunta es si el candidato se adaptará a las exigencias del puesto y en análisis de alimentos, si un alimento está bien clasificado de acuerdo con las múltiples características que posee. Normalmente, a falta de otra información, se toman las decisiones utilizando su propia experiencia o su intuición, para anticipar el comportamiento del individuo, a medida que van apareciendo los clientes o las muestras de alimentos. No obstante, cuando los problemas son más complejos y las consecuencias de tomar malas decisiones se hacen más costosas o graves, las decisiones basadas en la experiencia o en la intuición deben ser sustituidas por métodos más fiables y robustos y, especialmente, que controlen o minimicen la probabilidad de la mala decisión.

El análisis discriminante nos proporciona una metodología que permite realizar este tipo de clasificaciones. Es un método multivariado que facilita la clasificación de un grupo de objetos a través de p variables (llamadas discriminantes), en k grupos, definiéndose un modelo en términos de las variables discriminantes (ecuación discriminante), el cual permite clasificar observaciones futuras. La metodología clásica multivariante del análisis discriminante tiene un supuesto muy importante: las variables discriminantes tienen una distribución normal multivariada y a partir de allí se construye la ecuación discriminante. Esta es muy sensible a las observaciones atípicas, de tal forma, que puede cambiar

significativamente sin su presencia, generando clasificaciones falsas de nuevas observaciones, razón por la que, el análisis discrimínate robusto es una metodología de clasificación basada en la utilización de una submuestra de datos que no son atípicos y de un algoritmo de identificación secuencial de este grupo, el cual, finalmente, permite hacer la estimación de la ecuación discriminante. Esta metodología tiene también el supuesto de la distribución normal multivariada.

En el presente proyecto se analiza el problema de la discriminación de una población cuando tienen observaciones atípicas (*outliers*) aplicando la metodología de redes neuronales (clasificación robusta utilizando redes neuronales), y su objetivo es compararla con la metodología de discriminación robusta, analizando sus ventajas.

2. Planteamiento del problema

Desde que Fisher (1936) propuso la metodología de clasificación discriminante lineal, y los posteriores desarrollos de análisis discriminante cuadrático bayesiano (Geisser, 1964), y luego la versión robusta (Hubert y Van Driessen, 2004), uno de los problemas que se ha presentado es el del supuesto sobre el cual se basó el modelo y es el de normalidad.

La función discriminante cuando se hace el proceso de estimación de los parámetros requiere el supuesto de normal multivariada del conjunto de variables, sobre el cual esta se define. Y este supuesto es bastante restrictivo, debido a que, generalmente, no se cumple, y en otros casos, a pesar de que se cumple, es afectada por las observaciones atípicas (*outliers*). El problema tiene dos dimensiones: el no cumplimiento del supuesto de normalidad y la presencia de las observaciones atípicas.

Cuando se cumple el supuesto de normalidad, pero hay observaciones atípicas, existe la versión robusta; el problema se presenta cuando el supuesto de normalidad no se cumple y hay presencia de observaciones

atípicas; en este caso, el modelo robusto no es apropiado. Por lo anterior, se ha desarrollado una metodología para el análisis discriminante a través de redes neuronales, que no tiene el supuesto de normalidad de los datos, y se ha venido utilizando durante dos décadas. Aquí se plantea un modelo de discriminación a través de redes neuronales, que permita identificar la presencia de observaciones atípicas.

3. Estado del arte

3.1 Análisis discriminante

El problema del análisis discriminante se basa en la asignación de individuos a una de *m* poblaciones, con base en la caracteriza-ción, usando *p* variables, que se van a notar en un vector X, *p*-dimensional. Lo anterior, a través de la muestra aleatoria de vectores *p*-dimensionales X_{ik} los cuales se obtienen de poblaciones normales multivariadas *p*k, es decir:

$$\boldsymbol{X}_{ik}: \pi_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (i = 1, \ldots., n_k; \;\; k = 1, \ldots., m)$$

Donde n_k es el tamaño de la muestra en la población *k* para cada uno de los *m* grupos diferentes. También se asume que las matrices de covarianza son iguales ($\Sigma_1 = \ldots. = \Sigma_m = \Sigma$); , de esta forma, la probabilidad global de la mala clasificación es minimizada, por medio de la asignación de una nueva observación X a la población π_k la cual maximiza

$$d_k(X) = \frac{1}{2}(X - \boldsymbol{\mu})'\Sigma^{-1}(X - \boldsymbol{\mu}) + \log(\alpha_k) \qquad (k = 1, \ldots., m)$$

En este caso, α_k es la probabilidad a *priori* de que un individuo pertenezca a la población π_k. Los d_k (X) son los puntajes discriminantes, los cuales van a permitir decidir tomar la decisión de si la observación X pertenece a la población π_k *si* d_k *(X)*>d_j *(X)* para todo j=1,…,m. (Jhonson &y Wichern, 2008). Si las medias μ_k, con k=1,….., m, la matriz de covarianza ***S*** y las α_k

son desconocidas, que es el caso general, se requiere de una muestra de la población para poderlas estimar. El análisis discriminante cuadrático clásico (ADCC), utiliza las estimaciones de las medias muestrales de cada grupo X_j y la matriz de covarianza muestral S_k, como estimaciones de μ_k y Σ. Las probabilidades a *priori* α_k de las observaciones en cada grupo, se hacen a partir de $p_j=n_j/n$, donde n_j es el número de observaciones en el grupo j.

3.2 Estadística robusta

La estadística robusta estudia el comportamiento de los datos cuando existe una pequeña variación en los supuestos iníciales o cuando el modelo está contaminado por ciertas observaciones que tienen una malas influencias en los resultados (observaciones atípicas), proporcionando modelos erróneos o sesgos en la información suministrada.

La estadística robusta es necesaria e importante cuando se estiman modelos estadísticos para estudiar el comportamiento o la interrelación de las variables; dichas estimaciones se hacen a partir de muestras y se espera que el modelo estimado refleje de una manera fiel las interrelaciones "reales" de las variables en la población. La estimación de tales modelos, están dependiendo; así mismo, dependen totalmente de las muestras y de la forma como fueron tomadas, es decir, que se hayan tomado en condiciones similares y sin errores (Ortega, 2006). Así se pueden encontrar observaciones no deseables en la muestra, que perturban y proporcionan información errónea sobre la estructura de la mayoría de las observaciones; por tanto, se debe estudiar su concordancia con el resto de elementos de la muestra.

Cuando se tienen bases multivariadas de datos, estas posiblemente, pueden contener observaciones atípicas que no siempre son fácilmente identificables. En especial, cuando se analizan los datos de una forma global, pueden aparecer patrones complejos debido a la estructura misma de los datos, a las correlaciones entre las variables y a comportamientos ocultos, que solo pueden ser detectados cuando se hace un análisis más racional y profundo.

Johnson (1992), define: un atípico es una observación en un conjunto de datos que parece no coherente con el resto del conjunto de datos. De acuerdo con lo anterior se puede concluir que ”los atípicos no se ajustan a patrones que tenemos en nuestra mente, que se sale del campo de nuestra comprensión o de la estructura del comportamiento común (de la mayoría) de los datos, en donde subyace un comportamiento aleatorio”.

Existen distintos tipos de observaciones atípicas. Knorr (1998), clasifica los atípicos en simples y estructurales. Los atípicos simples surgen cuando la causa que los generan aparece en uno o más atributos de los individuos (observaciones); y los estructurales, cuando la causa que los generan emergen en todos los atributos o afectan a las relaciones entre las variables. En general, las observaciones atípicas multivariadas son aquellas que se consideran “extrañas”, no solo por el valor que toman en una determinada variable, sino por el conjunto de ellas. Son mucho más difíciles de identificar que los atípicos univariados, dado que no pueden considerarse como “valores extremos” como se hace en el caso univariado (Ganadesikan & Kattenring, 1972; Campbell, 1978).

Un aspecto muy importante de la estadística robusta es la detección e identificación de las observaciones atípicas multivariadas, la cual solo es posible, cuando se consideran las relaciones entre las variables dentro del análisis; esto hace que puedan existir muchas formas de aparecer atípicos. Cuando se analizan datos multivariantes, especialmente cuando se desea identificar las observaciones atípicas, el interés se centra en la medida de localización (centro) y la dispersión de los datos. La localización se describe mediante el vector de medias, el cual representa un punto en el espacio multidimensional; y la dispersión (o forma), por la matriz de covarianzas. La presencia de atípicos hace que las medidas de localización y de dispersión se sesguen, afectando la posibilidad de detección misma, debido a la eventualidad de que el sesgo se genere en la dirección de un solo atípico o de un grupo de atípicos, ocultándolos de atípicos dentro de la dispersión misma.

3.3 Estimadores para la detección de atípicos multivariados

Para detectar la presencia de atípicos multivariados se requiere hacer una estimación de las medidas de localización y dispersión de los datos multivariados. Los estimadores deben tener múltiples propiedades., Maronna et al.s (2006) presentan todas las propiedades que debe tener un buen estimador robusto multivariado. Aquí, se consideran solo las propiedades relacionadas con la detección y no las propiedades que se usan para otros objetivos. Se analizan, entonces, las siguientes cuatro propiedades:

- **Eficiencia y consistencia:** la eficiencia de un estimador está asociada con la varianza más pequeña. La consistencia de un estimador está asociada al comportamiento asintótico cuando la muestra crece;, esta propiedad se usa, particularmente, en el estudio de los estimadores robustos. La consistencia es muy importante cuando, por la complejidad de la definición de los estimadores, solo es posible estudiar su comportamiento asintótico.

- **Equivarianza**: esta propiedad se relaciona con la invarianza que tienen los estimadores a transformaciones de escala y de posición. Esta permite que al hacer una transformación de los datos, los estimadores no se afecten y sigan siendo equivalentes a los originales.

- **Punto de ruptura**: esta propiedad se relaciona con la fracción de puntos en la muestra que contaminan eal estimador, con observaciones atípicas., esta propiedad es similar a la del sesgo que mide la desviación causada en el estimador por la contaminación de la muestra; lo ideal es que el estimador capture la máxima desviación.

- **Eficiencia computacional:** esta propiedad se relaciona con la capacidad del estimador de ser calculado en el menor número de cálculos y en el menor tiempo de computacional.

Se han desarrollado múltiples estimadores robustos que contienen estas propiedades, los cuales han sido mejorados, especialmente en su

eficiencia computacional, que es quizás la principal preocupación actual, de todos los estimadores desarrollados. Se van a analizar tres:

- **Estimador de mínimo volumen elipsoidal (MVE):** este estimador fue propuesto por Rousseeuw (1984), y ha sido extensamente estudiado para la detección de atípicos multivariados. Se busca el elipsoide de mínimo volumen que cubre un subconjunto de al menos *h* datos. El estimador MVE consiste de dos parámetros que son el centro y la covarianza de una submuestra de tamaño h, ($h \leq n$) y que minimiza el volumen del elipsoide definido por la matriz de covarianza de la submuestra. Formalmente, el estimador se define como:

$$MVE = \left(\bar{x}_j^*, S_j^*\right)$$

Con $J = \left\{h\,observaciones : Vol(S_j^*) \leq Vol(S_K^*)\ para\ todo\ K\ que\ \#(K) = h\right\}$. El valor de h es usualmente $h = \left[\frac{m+p+1}{2}\right]$, donde [□] es la función del entero más grande.

- **Estimador basado en el determinante de covarianza mínima (MCD):** un procedimiento de estimación alternativa de alta ruptura al estimador MVE es el estimador basado en el determinante de covarianza mínima (MCD), que fue también propuesto por Rousseeuw (1984), este se obtiene encontrando el conjunto medio que da el mínimo valor del determinante de la matriz de varianza-covarianza. Formalmente, se puede definir como el centro y la covarianza de una submuestra de tamaño que minimiza el determinante de la matriz de covarianza asociada a la submuestra. Formalmente, el estimador se define como:

$$MCD = \left(\bar{x}_j^*, S_j^*\right)$$

Donde $J = \left\{h\,observaciones : \left\|S_j^*\right\| \leq \left\|S_K^*\right\|\ para\ todo\ K\ que\ \#(K) = h\right\}$.

- **Estimador T^2 de Hotelling Robusto:** un estimador usado ampliamente en análisis multivariado es el basado en la distancia de Mahalanobis.

$$MD_i = \sqrt{(x_i - T(X))C(X)^{-1}(x_i - T(X))'}$$

Donde T (X) es la media aritmética del conjunto de datos X, y C (X) es la matriz de covarianzas muestral. La distancia MD_i, establece cuán lejos está X_i del centro de la nube de datos (que casi siempre es la media de X). Un grupo de atípicos afectará a T(X) y será inflada la C(X) en esa dirección. Por consiguiente, parece natural reemplazar a T(X) y C(X) en la fórmula inmediatamente anterior, por estimadores robustos, y así obtener un estimador de distancia robusto.

A partir de la distancia de Mahalanobis, *s* se puede definir un estimador que poseae las mismas propiedades de esta distancia y es el T^2; este estimador fue definido por Hotelling (1952), y ha sido ampliamente estudiado en el campo del control estadístico multivariado:

$$T_i^2 = (x_i - T(X))C(X)^{-1}(x_i - T(X))'$$

El estimador T^2 de Hotelling es equivalente al cuadrado de la distancia de Mahalanobis y se ha demostrado que es efectivo para la detección de atípicos simples de tamaño moderado.

Desde el punto de vista de una muestra multivariada, se puede definir el estadístico T^2 de Hotelling utilizando T(X) y C(X) y basado en estimadores como MEV y MCD. Vargas (2003), hizo un análisis comparativo del desempeño del estimador T^2 de Hotelling Robusto.

3.4 Análisis discriminante robusto

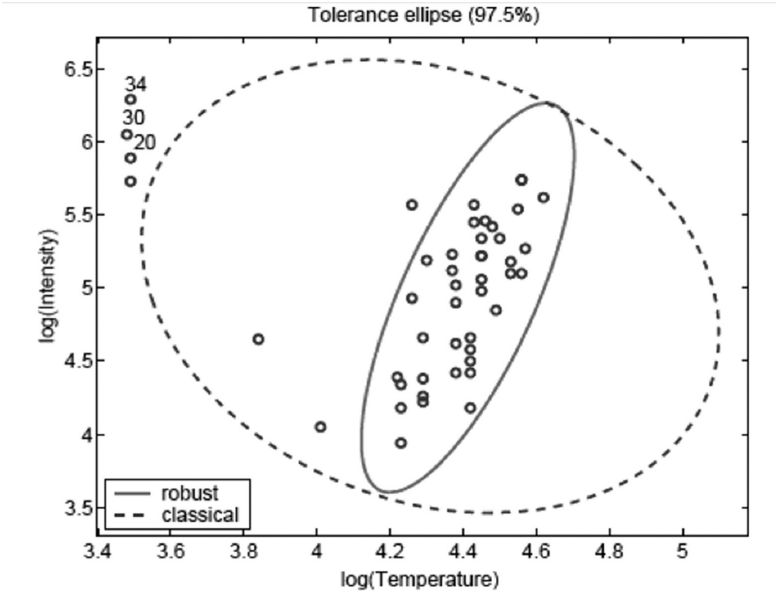
Para definir un estadístico de la función discriminante, se han propuesto múltiples estimadores (Todorov et al., 1994; Chork & Rousseeuw, 1992; Hawkins & McLachlam, 1997; He & Fung, 2000; Croux & Dehom, 2001). Aquí se va a utilizar el método basado en el estimador de determinante de covarianza mínima (MCD), en el cual se utiliza la distancia de Mahalanobis robusta:

$$MD_i^R = \sqrt{(x_i - \hat{\mu}_{MCD})\Sigma_{MCD}^{-1}(x_i - \hat{\mu}_{MCD})'}$$

Donde $\hat{\mu}_{MCD}$ y Σ_{MCD}^{-1} son los estimadores de localización y dispersión obtenidos a partir del estimador MCD. Este estimador propuesto por Hubert y Van Driessen (2004), donde utiliza el algoritmo FAST-MCD, y que garantiza la eficiencia computacional.

Se muestra el gráfico de distancias de Mahalanobis robusta y no robusta, que se reflejan en las elipses de confianza del 95%, en las que se aprecia que cuando se incluyen las observaciones atípicas, da una elipse muy distinta tanto en dispersión como de orientación, que cuando se excluyen las observaciones atípicas (distancia de Mahalanobis robusta) /Figura 1).

Figura 1. Elipse robusta y no robusta



El estadístico que permite estimar la función discriminante, que se denominará estimador discriminante cuadrático robusto (RQDR), está dado por:

$$\hat{d}_j^{RQ}(\boldsymbol{X}) = -\frac{1}{2} ln|\hat{\Sigma}_{j,MCD}| - \frac{1}{2}(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_{j,MCD})'\hat{\Sigma}_{j,MCD}(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_{j,MCD}) + \ln(\hat{\alpha}_j^R)$$

Con esta función discriminante se puede establecer si una nueva observación X pertenece a la población π_k si $\hat{d}_k^{RQ}(\boldsymbol{x}) > \hat{d}_j^{RQ}(\boldsymbol{x})$ para todo j = 1, ..., m, con j_k.

Para el caso lineal, se requiere estimar la matriz de covarianza común ***S***. Para esto, se han desarrollado tres enfoques: , el primer método consiste en ponderar las matrices de covarianza robusta de cada grupo, mientras que el segundo método pondera las observaciones (He &y Fung, 2000). Un tercer método, está basado en el algoritmo rápido para aproximar el determinante dentro del grupo (MWCD) desarrollado por Hawkins y McLachlan (1997).

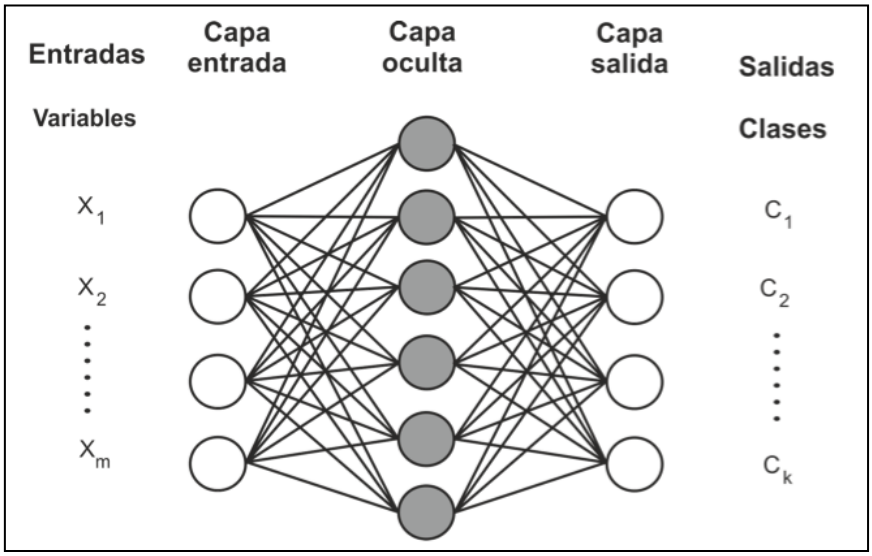
3.5 Análisis discriminante usando redes neuronales

La aplicación de redes neuronales a los estudios de análisis discriminante se ha desarrollado como un método alternativo y complementario debido a que se ha demostrado que el análisis discriminante clásico se ha mostrado que es muy restrictivo en sus supuestos (García & Moro, 1997) y es afectado por las observaciones atípicas; esto supone que la distribución de las variables discriminantes es normal multivariada, presume relaciones de linealidad y es sensible a las observaciones perdidas.

El planteamiento de usar las redes neuronales, nace especialmente de aplicaciones desarrolladas en los años 80 para el reconocimiento de imágenes, época en la que se propusieron métodos de clasificación y discriminación de datos (Waibel et al., 1989; Lefebvre et al., 1990; Bellustin el al., 1990; Van Allen et al., 1990; Maravall el al.,1991; Gemello y Mana, 1991). Adicionalmente, se desarrollaron aplicaciones dentro de las ciencias biológicas que fueron la génesis de los métodos modernos para la investigación del genoma (Qian & Sejnowski, 1988; Andreassen et al., 1990) y otras aplicaciones de modelos hidrobiológicos (Leck et al., 1994). A partir de este momento se desarrollaron un sinnúmero de aplicaciones en entornos financieros (Hawley et al. 1990; Tam et al., 1992; Altman et al., 1994)

Las Redes Neuronales Artificiales (RNA), son técnicas de procesamiento de datos con una estructura y unas funciones que imitan el sistema neuronal humano y más concretamente el funcionamiento del cerebro. De una forma simplificada, la RNA se puede conceptualizar como la conexión de diversas unidades básicas de procesamiento de información, llamadas neuronas, a través de tres tipos de capas: de entrada, ocultas y de salida. (Figura 2).

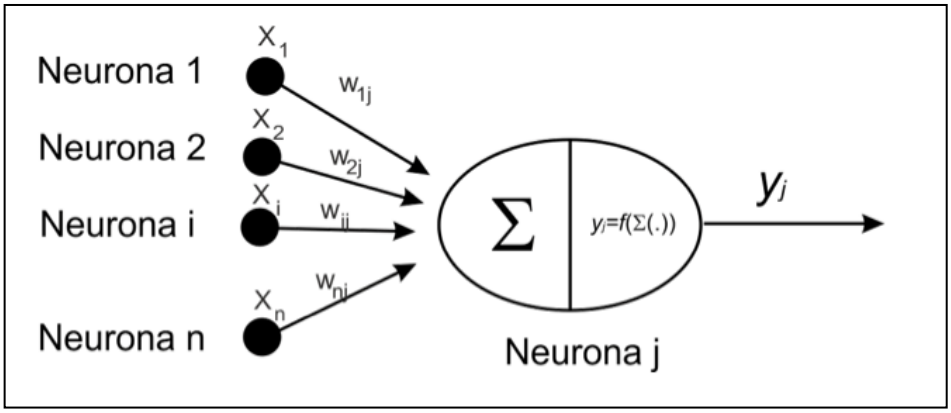
Figura 2. Red neuronal discriminante



Fuente. Elaboración propia de los autores.

A continuación se ilustra el funcionamiento conceptual de una red neuronal (Figura 3). Una neurona recibe información que le suministran las neuronas de la capa precedente, que la cual vendrá dada por la combinación lineal de la información proveniente de cada una, ponderada por cada uno de los coeficientes moduladores (en cada capa), denominados pesos sinápticos. De esta misma forma, una neurona proporciona una salida hacia cada una de las neuronas de las capas posteriores que se hallan a través de una función de transferencia o de activación. Existen varias funciones de activación (Anderson, 2000). Las X_i representan la salida de la *i-ésima* neurona precedente, w_{ii} es el peso sináptico que liga cada neurona con la neurona *j-ésima* y $f(\cdot)$ es la función de activación, que determina la salida de la j-ésima neurona (y_j) hacia el resto de neuronas.

Figura 3. Estructura de conexión de una RNA



Fuente. Elaboración propia de los autores.

Es importante señalar que existe una gran variedad de RNA, de pendiendo de la estructura (arquitectura). La forma que determinan los pesos sinápticos, se conoce como entrenamiento. El tipo de redes neuronales que usaremos se usarán en este trabajo es el Perceptron Multicapa. Un Perceptron Multicapa (PMC) es una red neuronal artificial que está basada en la capacidad de generalizar la información suministrada en una muestra compuesta por entradas y salidas, y asociar a una entrada, no suministrada anteriormente, una salida, de acuerdo con la estructura y las relaciones de información de la muestra. Este tipo de redes ajustan los pesos mediante entrenamiento supervisado (Anderson, 2000). A la red se le indica, para cada entrada suministrada, cuál es la salida deseada, buscando minimizar una función de error (generalmente, error cuadrático), entre la salida real y la salida que estima la red.

Un PMC consta de una entrada, a la cual se le suministran las observaciones de las variables independientes. Esta capa está conectada a un grupo de capas intermedias que se les llaman capas ocultas,; las cuales están unidas con la capa de salida, que proporciona el valor estimado para las variables explicadas. La entrada a una neurona *j*, net_j viene dada por:

$$net_j = \sum_i w_{ij} x_i + \theta_j$$

Donde t_j es el término conocido como humbral (sesgo) de la neurona j. Así mismo, para la salida de la neurona *j*, y la función de transferencia más utilizada es la función sigmoidal:

$$y_j = \frac{1}{1+e^{-net_j}}$$

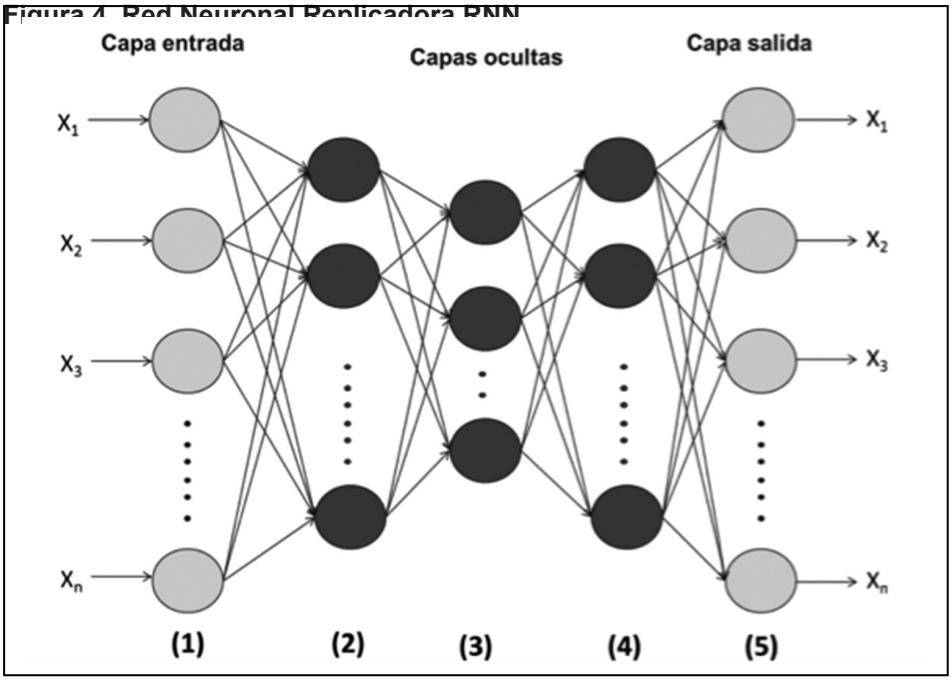
Los pesos y los umbrales que deben minimizar el error cuadrático de las salidas que proporciona la red, se hallan mediante el algoritmo *backpropagation* descrito por Rummelhart y McClelland (1986), basado en la optimización no lineal mediante el descenso del gradiente. Este consiste en un proceso iterativo en el que se muestran a la red las observaciones seleccionadas para su entrenamiento, y a partir del error que se ha cometidocometa en cada observación, se ajustan los pesos.

Respecto al número de capas ocultas, Martin y Sanz (1997), afirman que no es común diseñar redes con más de dos. La inclusión de una segunda capa oculta, mejora ligeramente la capacidad de aprendizaje de la red, pero aumenta el tiempo necesario de entrenamiento; por otro lado, el número de neuronas que debe tener la capa oculta (que determinará el número de parámetros por estimar), debe ser suficiente para que la red aprenda los ejemplos de la muestra de entrenamiento. Sin embargo, su número no debe ser excesivo, ya que puede producir un sobreaprendizaje que genere que, aunque la red ajuste muy bien la muestra de entrenamiento, no sea capaz de generalizar el conocimiento adquirido, proporcionando muy malas estimaciones cuando se le provee las entradas no suministradaos durante el aprendizaje. Desde el punto de vista de la clasificación y discriminación, la salida de la red suministra las clases que se obtienen desde el proceso del entrenamiento, las cuales caracterizan el proceso de discriminación.

3.6 Redes neuronales y la detección de atípicos

Los métodos multivariados robustos, en general, están soportados sobre el supuesto de normalidad multivariada de los datos. Se ha encontrado que este supuesto no siempre se cumple, y en ocasiones cuando se hacen la pruebas de ajuste de la distribución normal, esta es afectada directamente por la presencia de las observaciones atípicas; por lo tanto, frente a la presencia de estas observaciones aberradas, este supuesto es una limitante para la fiabilidad de los métodos multivariados.

Figura 4. Red Neuronal Replicadora RNN



Fuente. Elaboración propia de los autores.

Por esto, en ocasiones, es se hace necesario en ocasiones recurrir a métodos de detección de atípicos no paraméetricos. Hawkins et al. (2002), han propuesto un método de detección de atípicos en conjunto de observaciones atípicas a través de lo que se ha llamado un Redes Neuronales Replicadoraes (*Replicator Neural Network-RNN*). Una Red Neuronal Replicadora (*Replicator Neural Network–RNN*) es una red neuronal multicapa tipo perceptróon con tres capas ocultas entre las capas de entrada y la de salida. La función de la Red Neuronal Replicadora es reproducir el patrón de entrada de los datos en la salida, con el mínimo de error, a través del entrenamiento de los datos. Tanto las capas de entrada como las de salida tienen *n* unidades, correspondientes a las *n* características de los datos entrenados. El número de unidades de las capas ocultas se seleccionan experimentalmente, lo que permite minimizar los errores de reconstrucción. Hawkins (2000), hace un análisis de la heurística utilizada en la experimentación. (Figura 4).

La salida de la *i-ésima* unidad de la capa *k* se calcula usando la función de activación S_k (I_{ki}), donde I_{ki} es la suma ponderada de las unidades, definida como:

$$I_{ik} = \sum_{j=0}^{L_{k-1}} w_{kij} Z_{(k-1)j}$$

Z_{kj} es la salida de la j-ésima unidad de la k-ésima capa y L_k es el núumero de unidades de la k-éesima capa. Para cada capa de la red se utiliza una función de activación distinta; de esta forma, la función de activación de la primera y última (2-4) capas ocultas es de la forma S_k (I_{ki})=tanh (a_k I_{ki}), donde a_k es el parámetro de sintonización. Para la capa oculta intermedia (3), la función de activación es de tipo escalonado, de la forma siguiente:

$$S_k(I_{ki}) = \frac{1}{2} + \frac{1}{2(k-1)} \sum_{j=1}^{N-1} tanh\left[a_3\left(I_{ki} - \frac{1}{N}\right)\right]$$

Donde *N* es el número de paso o niveles de activación y a_3 es el parámetro que controla la tasa de transición de un nivel a otro.

Entrenamiento: para la capa de salida final se selecciona una función de activación de tipo sigmoide.

$$S_5(I_{ki}) = \frac{1}{1 + e^{-a_5 I_{ki}}}$$

Se utiliza una tasa de aprendizaje adaptativo para el entrenamiento de la red, en cada iteración *l*, los pesos de actualización de la red son de la forma:

$$w_{ij}^{l+1} = w_{ij}^{l} + \alpha_{l+1} \Delta w_{ij}^{l+1}$$

De la forma como se hace la detección de la atipicidad de una observación, aplica el llamado Factor Atipicidad (FA). Se define el Factor Atipicidad del dato *i-ésimo* como FA_i , el error promedio de reconstrucción de todas las características (unidades).

$$FA_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - o_{ij})^2$$

El FA se evalúa para todo los datos usando la RNN entrenada. Hawkins et al. (2002), Wiliams et al. (2002) y Bakar et a.l (2006) discuten más en detalle sobre la metodología de identificación de atípicos mediante Redes Neuronales Replicadoras.

En consecuencia, los replicadores neuronales usados para la identificación de observaciones atípicas, lo que hacen es intentar reproducir los patrones de las entradas en las salidas, durante el proceso de entrenamiento, los pesos de la RNA se ajustan para minimizar el error cuadrático medio (o error de reconstrucción promedio) para todos los patrones de entrenamiento. Entonces, los patrones comunes son los más propensos a ser bien reproducidos, por el entrenamiento de la RNA; así que aquellos patrones que representan atípicos, son los menos bien reproducidos por la RNA entrenada y tendrán un mayor error de reconstrucción. El error de reconstrucción se usa como medida de atipicidad de un dato.

3.7 Aplicación del método

Para realizar la aplicación del método de análisis discriminante robusto, se va a estudiar un caso donde en el que el análisis discriminante tiene una gran aplicación y es en: los estudios morfoméetricos de las abejas. Las abejas son un sector de alta productividad de un país; es una industria que tiene un alto crecimiento en Colombia. El nivel de producción del sector depende de la capacidad que tengan las colmenas de producir miel y otros productos complementarios. La productividad de la colmena depende directamente de la raza de abeja. En Colombia está difundida la abeja europea *Apis Mellifera (L)* que fue introducida en el siglo XVII; esta se adaptó bastante bien en los climas templados. A mediados de la década de los años 50 se trajeron de Brasil varias colmenas de la abeja *Apis Mellifera Scutellata* de origen africano (las cuales son de alta productividad de miel), con el fin de realizar cruces y producir híbridos de alta productividad. Estas abejas se caracterizan porque atacan frecuentemente.

Tabla 1. Variables medidas en el estudio morfométrico

Notacion	Variables
V1	Longitud del ala anterior derecha
V2	Ancho del ala anterior derecha
V3	Longitud del ala posterior derecha
V4	No. de garfios del ala posterior
V5	Longitud de la trompa
V6	Longitud de la tibia pata posterior
V7	Longitud fFémur pata posterior
V8	Ancho de cuarto tergito
V9	Longitud del cuarto tergito
V10	Ancho del cuarto esternito
V11	Ancho del cuarto esternito

Fuente. Elaboración propia de los autores.

Esto obliga a que se deba tener un control de este tipo de abejas, porque a pesar de que son muy productivas generan problemas de rechazo y exterminio; por lo tanto, se deben tener algunos criterios para determinar su grado de africanización. Las abejas africanizadas son muy similares a las otras, pero en algunas características morfométricas tienen algunas diferencias, que deben considerarse y evaluarse para poder controlarlas, anticipándose al problema.

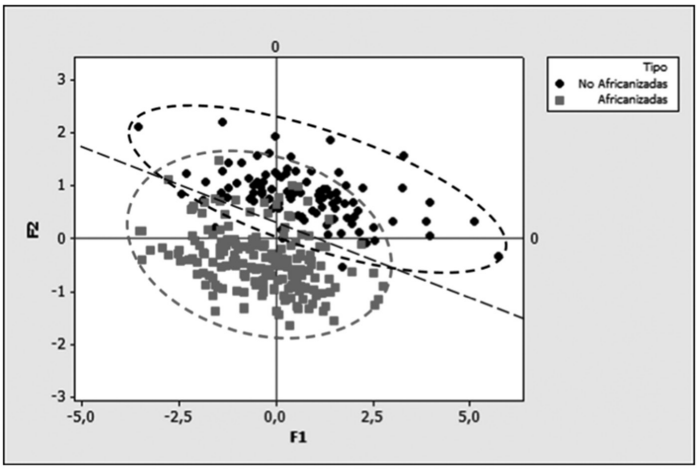
Son muchos los estudios en los que se ha aplicado el análisis discriminante para evaluar el grado de africanización o la caracterización genética para otros estudios; el primer estudio de análisis morfométrico con análisis discriminante data de 1948 (Altapov), que dio una metodología de medición morfométrica, la cual aún se utiliza.

3.8 Análisis discriminante clásico y robusto

Los datos del presente estudio son parte de una investigación desarrollada en la Universidad del Tolima (Salamanca et al., 2001), en la que se evaluaron muestras de abejas en colmenas de diferentes zonas de este Departamento. Se sabe que el grado de africanización está influenciado

por factores climáticos y medioambientales, así como el grado de hibridación de las abejas. Como parte del estudio, se tomaron muestras de abejas no africanizadas y africanizadas y se les midieron las variables que caracterizan la diferenciación entre razas de abejas, según Daly y& Balley (1978) (Tabla 1). En total se analizaron 390 abejas, de las cuales 110 eran africanizadas. El análisis discriminante en el que se incluyeron todas las observaciones se ilustra a acontinuación (Figura 5).

Figura 5. Análisis discriminante clásico



Posteriormente, se llevó a cabo el análisis robusto, suponiendo que tiene un comportamiento normal. Para realizar el análisis robustos, primero se hizo la estimación de la covarianza del mínimo determinante MCD, aplicando el algoritmo de convergencia rápida desarrollado por Rousseeuw, & Van Driessen (1999); se utilizó el paquete LIBRA de Matlab, especializado para análisis Robusto desarrollado por Verbone & Hubert (2004) de la Universidad Católica de Lovaina. Este módulo está desarrollado en *R* también en el paquete "rrcov".

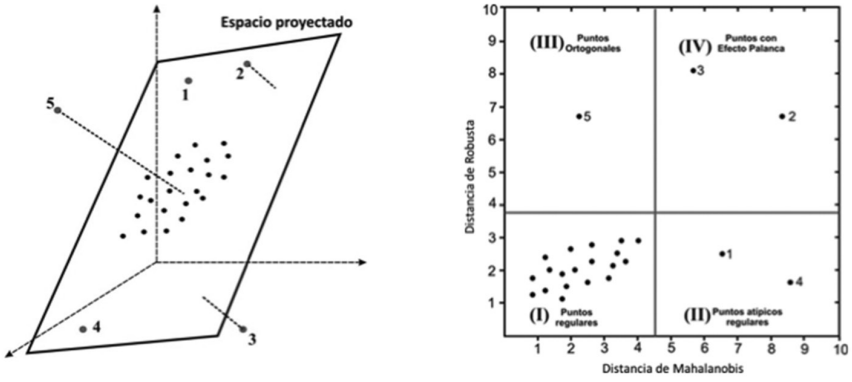
Se identificaron las observaciones atípicas a través del diagrama de distancias robustas, donde para cada observación se calculó la distancia de Mahalanobis y la distancia robusta, así:

$$MD_i = \sqrt{(X_i - \bar{X})^T \Sigma^{-1} (X_i - \bar{X})}$$

$$DR_i = \sqrt{(X_i - \hat{\mu}_{MCD})^T \Sigma_{MCD}^{-1} (X_i - \hat{\mu}_{MCD})}$$

Estas distancias se resumen en la gráfica de diagnóstico de atípicos, en la que para cada observación se muestra en un eje coordenado con las dos distancias (Figura 6).

Figura 6. Diagnóstico de observaciones atípicas



La figura de diagnóstico de observaciones atípicas, utiliza la proyección factorial en la que se distinguen cuatro tipos de observaciones atípicas:

- **Los puntos regulares** que forman un grupo homogéneo que está muy cerca al centro de inercia del espacio proyectado.

- **Los puntos atípicos regulares** que están dentro del espacio proyectado, pero lejos de las observaciones regulares; estas son como las observaciones 1 y 4.

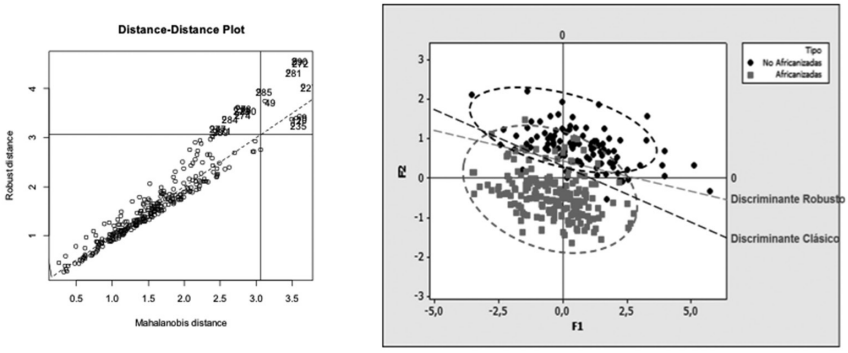
- **Los puntos atípicos ortogonales** que están a una distancia ortogonal muy grande del espacio proyectado; no obstante, su proyección cae dentro del conjunto de puntos regulares. La observación 5 es un punto atípico ortogonal.

- **Los puntos atípicos con efecto palanca** que tienen una distancia ortogonal grande y su proyección cae lejos del conjunto de puntos regulares, tales como las observaciones 2 y 3.

A partir de esta gráfica se pueden caracterizar las observaciones atípicas, e identificarlas. Para los datos de mediciones morfométricas, de las abejas aparecen con un comportamiento muy especial debido a la alta correlación entre dichas mediciones morfometricos. Allí se identificaron, en total, 12 observaciones de tipo palanca, que afectan directamente a la proyección y, por lo tanto, a la función discriminante.

Al estimar la covarianza de mínimo determinante, ΣMCD permite hacer la estimación robusta de la función discriminante, sin considerar las observaciones atípicas (Figura 7) en las que se comparan las dos funciones discriminantes y el sentido de las elipses de confianza que tienen menor variabilidad que con las observaciones atípicas.

Figura 7. Análisis discriminante robusto



Se observa que cuando se estima la función discriminante con la influencia de las observaciones atípicas, tienen una pendiente diferente haciendo que tengan un poder de discriminación distinta. Es posible que existan observaciones resulten mal discriminadas dependiendo de la función discriminante que se utilice. Por otro lado, se observa que la función discriminante robusta, tiene una menor dispersión.

3.9 Análisis discriminante robusto con redes neuronales

Para realizar el análisis discriminante con redes neuronales, primero se lleva a cabo el proceso de detección de atípicos, mediante una red neuronal replicadora.

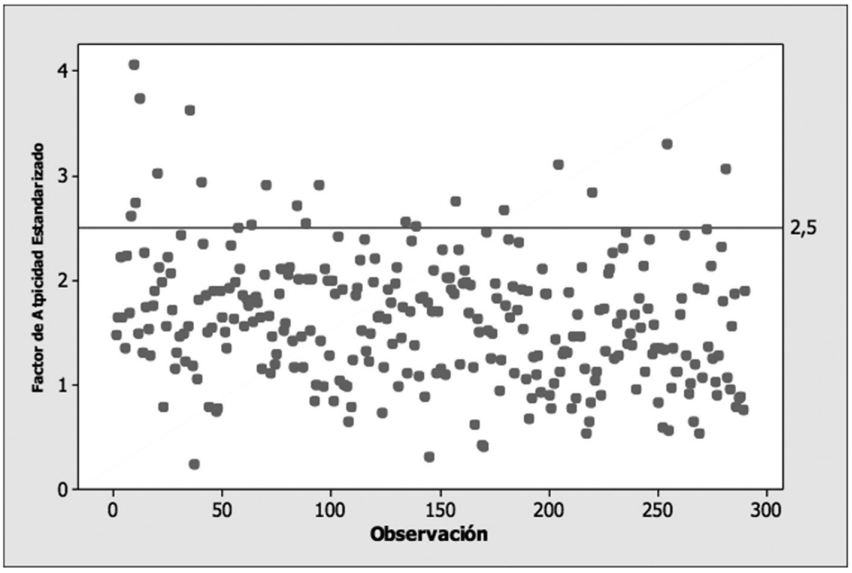
Para hacer la estimación de esta red, se tomó una muestra de la mitad de observaciones, seleccionada al azar, y se realizó el proceso de aprendizaje. Se utilizó una red de tres capas ocultas donde las observaciones de entrada son las mismas de salida. La metodología de estimación de la red es *backpropagation*, y se utilizó el módulo de herramientas de Redes Neuronales de Matlab.

Una vez se obtuvo la red aprendida se utilizó para estimar el FA, factor de atipicidad de todas las observaciones $FA_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - o_{ij})^2$. Para poderla representar gráficamente y hacerla comparable con el criterio de Mahalanobis, se estandarizó.

Se muestran los factores de atipicidad estandarizados de todas las observaciones (Figura 8). Se puede apreciar que hay observaciones a las que la red neuronal replicadora aporta un factor grande; para tener un valor de referencia con el cual se pueda tomar decisiones, se utiliza el mismo criterio de discriminación de la distancia de Mahalanobis en la identificación de atípicos de Análisis de Componentes Principales (Lebart et al., 2000). Este es un punto abierto para estudiar y analizar más en profundidad. Allí se detectaron los mismos 12 identificados por el análisis robusto usando la distancia robusta, y además aparecen 6 más que están muy cerca, que podrían interpretarse como observaciones atípicas enmascaradas.

A continuación se estimó una red neuronal clasificadora tipo peceptron doble capa, donde las entradas son las observaciones y las salidas son las categorías de observaciones (raza de abeja). Esta red no tiene complicación; es el tipo de redes neuronales que se han venido estimando desde que se desarrollaron las redes neuronales. Se observa que las redes discriminan bastante bien al detalle, aunque hay muy pocas observaciones que pasan por el umbral y pueden estar mal discriminadas.

Figura 8. Análisis del factor de atipicidad para la morfología



4. Conclusiones

En este trabajo se planteaa cómo integrar la clasificación o discriminación de observaciones usando redes neuronales (que es bastante usada) con la metodología de identificación de observaciones atípicas mediante redes neuronales replicadoras. Se muestra cómo a través de la herramienta de Factor de Atipicidad (F_{ai}) se puede identificar la distancia robusta; la diferencia radica en que en las redes neuronales, se hace en forma algorítmica y no requiere de supuestos de las distribuciones subyacentes.

Se muestra que con las observaciones atípicas, la función de discriminación cambia sustancialmente tanto en forma clásica-Robusta como con redes neuronales.

Aparecen varios interrogantes, que se van a estudiar, en futuras investigaciones. También se debe entrar a sistematizar en forma metodológica el paso a paso para aplicar esta metodología.

Agradecemos doctor Guillermo Salamanca por facilitarnos y dar permiso de utilizar los datos de su investigación, con el fin de ilustrar la metodología robusta con redes neuronales.

Finalmente, se puede observar, que al comparar la metodología de discriminante robusto con la metodología robusta usando redes neuronales, arrojan resultados muy similares. Se plantea hacer un análisis más en detalle utilizando el método *bootstrap* para hacer evaluaciones, estadísticas de la metodología.

5. Referencias bibliográficas

Altman, Ei.; Marco, G.; Varetto F (1994) *"Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks" Journal and Banking and Finance*, 18. P: 505-529

Andreassen, H., Bohr, H., Bohr, J., Brunak, S., Bugge, T., Cotterill, M.J., Jacobsen, C., Kusk, P., Lautrup, B., Petersen, S.B., Saermark, T. and Ulrich, K.,(1990). *"Analysis of the secondary structure of the human immunodeficiency virus (HIV) proteins p17, gp20, and gp41 by computer modelling based on neural network methods"*. Journal. Acq. Immun. Defic. Syndrome, 3: pp. 615–622.

Altman, Ei.; Marco, G.; Varetto F (1994) *"Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks"* Journal and Banking and Finance, 18. P: 505-529

Belliustin, N.S., Kuznetsov, S.O., Nuidel, I.V. and Yakhno, V.G., (1991). *"Neural networks with close nonlocal coupling for analysing composite image"*. Neurocomputing, 3: 231– 246.

Campbell, N.A. (1978) "*The influence function as an aid in outlier detection in discriminant analysis"*. Applied Statistics, 27, pp. 251-258.

Campbell, N.A. (1980), *"Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation*," Applied Statistics, 29, 231–237.

Chork, C. and Rousseeuw, P.J. (1992). "*Integrating a high breakdown option into discriminant analysis in exploration geochemistry"*, Journal of Geochemical Exploration, 43, 191–203

Croux, C. and Dehon, C. (2001). *"Robust linear discriminant analysis usings estimators"*, The Canadian Journal of Statistics, 29, 473–492

Fisher, R.A.,(1936). *The use of multiple measurements in taxonomic problems.* Annual Eugenics, 7, Part II, pp. 179-188. 1936.

García de Ceca J.L y Moro J (1997) *"Comparing Neural Networks and Multivariate Discriminat Analysis in the selection of new crop varieties" First European Conference for Information Technology in Agriculture,* Copenhagen, 15-18 June.

Gemello, R. and Mana, F., 1991. *"A neural approach to speaker independant isoled word recognition in an uncontrolled environment".* In: Proc. Int. Neural Networks Conf., Paris July 9–13, 1990, Kluwer Academic Publishers, Dordrecht, Vol. 1, pp. 35–37.

Geisser S. (1964). *Posterior odds for multivariate normal distributions. Journal of the Royal Society Series B Methodological*, 26:69–76.

Hawkins S., H. X. He, G. J. Williams, and R. A. Baxter (2002). *"Outlier detection using replicator neural networks"*. In Proceedings of the Fifth International Conference and Data Warehousing and Knowledge Discovery (DaWaK02).

Hawkins, D.M. and McLachlan, G. (1997). *"High-breakdown linear discriminant analysis"*, Journal of the American Statistical Association, 92, 136–143.

Hawley D.D., Jhonson J.D. y Raina D. (1990) *"Artificial Neural Sustems: a new tool for financial decisión-making"* Financial Analysis Journal. November-December. P: 63-72.

Hecht-Nielsen R. (1995). *"Replicator neural networks for universal optimal source coding".* Science, 269(1860-1863).

He, X. and Fung, W. (2000). *"High breakdown estimation for multiple populations with applications to discriminant analysis"*, Journal of Multivariate Analysis, 72, pp: 151–162.

Hubert, M. and Van Driessen, K. (2004). *"Fast and robust discriminant analysis"*, Computational Statistics and Data Analysis, 45, pp: 301–320.

Johnson R. A. Wichern D. W.(2008) *Applied Multivariate Statistical Analysis.* Prentice Hall

Johnson (1992). *Applied Multivariate Statistical Analysis*. Prentice Hall

Lefebvre, T., Nicolas, J.M. and Dagoul, P., (1990). *"Numerical to symbolical conversion for acoustic signal classification using a two-stage neural architecture"*. In: Proc. Int. Neural Networks Conf., Paris July 9–13, 1990, Kluwer Academic Publishers, Dordrecht, Vol. 1, pp. 119–122.

Lek, S., Lauga, J., Giraudel, J.L., Baran, P. and Delacoste, M., (1994). *"Example of an application of formal neural networks in the environmental sciences".* In: V. Bretagnolle and F. Beninel (Editors), Proc. Int. Meeting 'Ecology and statistical methods', C.N.R.S., Chize´, pp. 73–82.

Maronna R., Doug M. y Yohai V. (2006) *Robust Statistics*. Wiley

Maravall, D., Rfos, J., Pérez-Castellano, M., Carpintero, A. and Gımez-Calcerrada, J., (1991). *"Comparison of neural networks and conventional techniques for automatic recognition of a multilingual speech database".* In: A. Prieto (Editor), Artificial Neural Networks, Proc. Int. Workshop IWANN'91, Granada (SPAIN), Sep-91, pp. 377–384.

Maronna R., Doug M. y Yohai V. (2006) *Robust Statistics*. Wiley

Ortega J. F. (2006) *Notas sobre estadística Robusta. Universidad Castilla la Mancha.* España

Qian, N. and Sejnowski, T.J., (1988). *"Prediction of the secondary structure of globular proteins using neural networks models"*. Journal Mol. Biol., 202: 865–884.

Rousseeuw, P. J. (1984). *"Least median of squares regression"*. Journal of the American Statistical Association, 79(388):pp: 871—880

Rummelhart D. y McClelland J. (1986) *"Parallel distributed processing"* Cambridge (Massachussets) MIT press.

Salamanca G. Vargas E. y Perez E. C. (2002) *Estudio morfométrico y sistemático del Grado de Africanización de la Abeja Apis mellifera en algunas zonas del departamento de Boyacá*. http://www.beekeeping.com/articulos/salamanca/africanizacion_boyaca.htm

Tam K.Y., Kiang M. Y. (1992) *"Predicting bank failures: a neural networks approach"* Management Science 38, 7 p: 926-947.

Todorov, V.; Neykov, N. and Neytchev, P. (1994).*"Robust two-group dis crimination by bounded influence regression"*, Computational Statistics and Data Analysis, 17, pp: 289–302.

Van Allen, E., Menon, M.M. and Dicaprio, N., (1990). *"A modular architecture for object recongnition using neural networks*". In: Proc. Int. Neural Networks Confer., Paris July 9–13, 1990, Kluwer Academic Publishers, Dordrecht, Vol. 1, pp. 35–37.

Waibel, A., Hanazawa, T. Hinton, G. Shikano, K. and Lang, K.J., 1989. *Phoneme recognition using time-delay neural networks*. IEEE Trans. Acoust. Speech Signal Proc., 37(3): 328–339.