Complejidad de los datos en las Redes Neuronales Artificiales: Estado de la Cuestión

P. Toribio[1], B. G. Rodriguez[1], R. Alejo[2]

[1] Centro Universitario UAEM Atlacomulco, Universidad Autónoma del Estado de México, Km. 60 Carretera Toluca-Atlacomulco (México).
[2] Dept Llenguatges I Sistemes Informátics, Universitat Jaume I, Av. Sos Baynat S/N, 12071 Castelló de la Plana (Spain).

Resumen. En la actualidad, las redes neuronales artificiales son ampliamente utilizadas para tareas de reconocimiento de patrones, minería de datos y aprendizaje automático. Lamentablemente, su aplicación a problemas reales todavía está restringida por algunas debilidades como lentitud en el proceso de aprendizaje y la pobre capacidad de generalizar que han mostrado en múltiples aplicaciones prácticas.

En este trabajo se estudian las principales características de los datos de entrenamiento que tienen un efecto negativo en el proceso de aprendizaje de la red neuronal, por ejemplo: el ruido, el solapamiento de clases, desbalance de clases, principalmente. Actualmente, la complejidad de los datos ha atraído la atención de numerosos investigadores. Se ha observado que la capacidad de generalización y la convergencia de la red dependen en gran medida del nivel de complejidad en los datos de entrenamiento

Palabras clave: Redes Neuronales Artificiales, costo-sensitivo, over-sampling, under-sampling, Wilson.

1 Introducción.

Hoy en día el área de reconocimiento de patrones, aprendizaje automático y minería de datos han despertado el interés de muchos investigadores, ya que, se han convertido en herramientas para automatizar procesos en diferentes áreas como la medicina, la industria, la biología, etc. Por ejemplo, la minería de datos ha surgido como alternativa para la extracción del conocimiento de grandes cantidades de datos con la finalidad de tomar decisiones o comprender algún fenómeno social o de la naturaleza.

Actualmente, las Redes Neuronales Artificiales (RNA) son muy utilizadas como clasificadores en tareas de minería de datos, aprendizaje automático y reconocimiento

M.A Cruz-Chávez, J.C Zavala Díaz(Eds):CICos2009, ISBN:978-607-00-1970-8, pp. 229 - 235, 2009.

de patrones. El Perceptron Multicapa (PM) ha sido utilizado en la interpretación de imágenes de lectura remota [1] y como aproximador universal, principalmente. Sin embargo, aún se conoce poco de estos modelos, lo que trae como consecuencia, lentitud en el proceso de aprendizaje y pobre capacidad de generalización. Varios investigadores han coincidido en que la precisión de estos clasificadores depende de la calidad de los datos [2, 3, 4, 16], porque las RNA fueron diseñadas para aprender mediante bases de datos con clases balanceadas y también con datos que puedan ser linealmente separables, o por lo menos que para este tipo de clasificador sea fácil transformar el espacio no linealmente separable a uno linealmente separable, por lo tanto todos aquellos factores o características del conjunto de datos que impiden que una RNA aprenda correctamente reciben el nombre de complejidad de los datos [2, 3, 5, 6, 7].

Cada investigador define de forma diferente o engloba diferentes factores en complejidad de los datos, por ejemplo, Barandela en [3, 4] dice que el conjunto de datos para entrenar la red contiene patrones atípicos y ruidosos que la afectan negativamente, Visa *et al.* en [2] tratan varios factores como lo es el desbalance (el problema más grave en redes neuronales) [5, 6, 7, 8], el tamaño de la muestra de entrenamiento y el solapamiento de clases (ver Fig. 1).

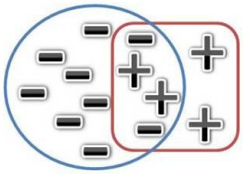

Fig. 1. Solapamiento de clases

2 Complejidad de los datos.

Como describimos en la sección anterior la complejidad de los datos son aquellos factores que disminuyen la calidad de la muestra de entrenamiento (ME) y que por consiguiente disminuyen la precisión del clasificador. Estos factores se pueden definir como sigue:

- **Ruido**: Son datos con errores, originados en su medición o registro (mal etiquetados, es decir, cuando no pertenecen a la clase donde fueron colocados).
- **Patrones Atípicos**: Son excepciones a la regla, es decir, que aunque han sido identificados correctamente estos son muy diferentes al resto de los patrones de la misma clase.
- **Solapamiento**: es cuando dos o más clases se encuentran interceptadas entre sí compartiendo elementos en común (ver Fig. 1).

- **Desbalance de clases**: sucede cuando una clase tiene más patrones que las otras, es decir, cuando una clase es demasiado pequeña respecto a las demás.
- **Tamaño de ME**: La cantidad de patrones es directamente proporcional al tiempo que tarda en aprender una RNA todo el conjunto de estos, debido a esto, entre más grande sea la muestra más tiempo consumirá la red en aprenderlos, agregando la carga computacional asociado a este proceso. Algunos autores [8] no solo asocian el tamaño a la cantidad de patrones si no también a la cantidad de atributos de cada patrón.
- **Distribución de la los datos**: la Fig. 2 muestra dos casos en donde se observa cómo están distribuidos los datos. En la Fig. 2a se ve una distribución que para una red neuronal seria la óptima. Aquí tomaremos a la linealidad dentro de la distribución de los datos, pues la primera depende de la segunda.

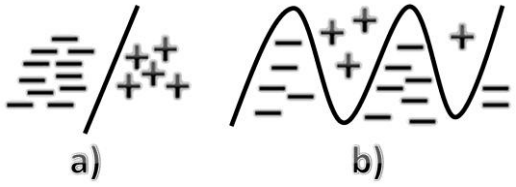

Fig. 2. Distribución de los datos, a) caso óptimo, b) caso no óptimo.

3 Métodos para tratar la complejidad de los datos

Particularmente, el problema más grave y al cual la mayoría de los investigadores se han enfocado es el desbalance de clases [2, 5, 6, 8, 10], y es que en la mayoría de los problemas reales los datos estas desequilibrados, por ejemplo, en los fraudes por vía telefónica (son mucho menos que las llamadas telefónicas normales) o al tratar con enfermedades como el SIDA donde hay menos personas contagiadas en comparación con las sanas. Se han propuesto varios métodos para corregir el problema del desbalance en los datos, entre los más conocidos son los métodos de muestreo [2, 10], basados en costo-sensitivo [11], los cuales se describen en la sección 3.1.

Para tratar el solapamiento de clases, patrones atípicos y ruido se utilizan métodos como la edición de Wilson, basados en la técnica del vecino más cercano (*nearest neighbor rule, NN*) y sus variantes, por ejemplo el *1-NN*, *k-NN* y *CNN*. También se han utilizados técnicas basadas en el criterio de vecindad como el Grafo de Gabriel (*GG*, *Gabriel Graph*) y Grafo de la vecindad relativa (*RNG*, *Relative Neighbourhood Graph*) [12], las cuales se presentan en un panorama más amplio en la sección 3.2. Cabe señalar que tanto estos métodos como los de muestreo para combatir el desbalance modifican la distribución de los datos, así como el tamaño de la ME, de esto podemos plantearnos la siguiente pregunta, ¿qué tan conveniente es modificar el tamaño y la distribución de los datos?

3.1 Métodos de muestreo (re-balancear la ME) y Costo-Sensitivo.

Cuando hablamos de corregir el desbalance de datos por muestreo, nos referimos principalmente al *over-sampling*, *under-sampling* [2, 10]. La función principal del *over-sampling* es añadir a la clase minoritaria copias de los patrones de esta misma clase hasta equilibrar las clases de manera aleatoria, aunque Haibo *et al.* en [10] nos dicen que la desventaja de usar este método es que se puede llegar a un sobre-ajuste [13]. De forma contraria el *under-sampling*, elimina patrones de manera aleatoria de la clase mayoritaria hasta balancear la ME. Además se ha popularizado *SMOTE*, que es una variación del *over-sampling*, la diferencia consiste en el hecho de que ya no se duplican patrones aleatoriamente, si no se crean patrones sintéticos mediante la interpolación [2]. Algunos autores, rechazan estos métodos porque el *under-sampling* involucra pérdida de información (por descartar datos potencialmente útiles), y el *over-sampling* incrementa el tamaño de ME sin ningún aumento de información, por lo que algunos expertos sugieren que las investigaciones se enfoquen a los algoritmos de aprendizaje [2].

Otro enfoque para combatir el desbalance de clases, es el costo-sensitivo (*cost-sensitive*) [11]. Este se basa en la afirmación de que el precio de cometer un error de clasificación debe ser distinto para cada clase [14]. En este método se implementa una matriz de costos para C clases (los valores dependen del problema que se esté tratando), donde $Cost[i,j](i,j,\epsilon\ [1 \dots C])$ denota el costo de una clasificación errónea de un patrón de clase *i*-ésima en la clase *j*-ésima, por ejemplo $Cost[1,2] = 1$ (ver Tabla 1). Además, durante los últimos años se ha popularizado el paradigma de aprendizaje en conjuntos [11, 18], es decir, se entrenan varios clasificadores y se combinan sus predicciones. En este paradigma, cada clasificador es un aprendiz, entonces cada uno de estos votan por una clase y la clase que recibió más votos es la que se retorna.

Tabla 1. Ejemplo de una Matriz de Costo *(Cost[i, j])* para una ME con tres clases.

		j		
		1	2	3
	1	0	1	8
i	2	1	0	9
	3	1	1	0

3.2 Métodos para eliminar el solapamiento, ruido, patrones atípicos y solapamiento de clases.

En [15,16] se habla del uso de técnicas basadas en la regla del vecino más cercano, para la eliminación de ruido en la ME, mejorando el rendimiento del clasificador. En la mayoría de las investigaciones se tratan en conjunto al solapamiento, ruido y patrones atípicos (englobaremos estos tres factores como limpieza de la ME). Los algoritmos más utilizados para limpiar la ME son los de Edición que consisten en

descartar prototipos (patrones) que se encuentren en la región correspondiente a una clase distinta a la suya, es decir, prototipos cuya probabilidad de pertenencia a su clase se vea superada por la probabilidad de pertenencia a alguna otra clase. La Edición de Wilson [17], propone eliminar los elementos atípicos de la muestra de entrenamiento mediante la aplicación del procedimiento *k-NN* (*k* vecinos más cercanos), dado un patrón de prueba de la ME se buscan sus *k*-vecinos más cercanos y si la mayoría estos no pertenecen a la misma clase, el patrón de prueba se elimina. La Edición de WilsonCN [18] es una modificación al algoritmo de Wilson. Este método se basa en usar, en vez de la regla *k-NN*, la regla *k-NCN* (regla de los *k*-Vecinos de Centroides más Cercanos) como clasificador central del método, debido a que la misma obtiene mejores resultados, sobre todo cuando el conjunto de entrenamiento es relativamente pequeño. Es importante destacar que este algoritmo presenta el inconveniente de tener un costo computacional superior a la variante original lo que limita su utilización en determinados problemas reales [18].

Además de métodos basados en la regla del vecino más cercano y su variantes los métodos basados en el criterio de vecindad han dado buenos resultados para la limpieza de la ME, aunque su impacto sobre la precisión del clasificador no ha alcanzado lo esperado. A continuación se describe con más detalle el funcionamiento de estos métodos.

La técnica *GG* (Grafo de Gabriel) [18] dice que para un conjunto *V* de *n* puntos (refiriéndose a un vector) $V = \{p_1, p_2, \ldots, p_n\}$, dos puntos p_i y p_j son vecinos de Gabriel si:

$$\text{dist}^2(p_i, p_j) < \text{dist}^2(p_i, p_k) + \text{dist}^2(p_k, p_j) \qquad \forall k \neq i, j \qquad (\mathbf{1})$$

Uniendo todos los vecinos de Gabriel en pares mediante una arista se obtiene el grafo de Gabriel. En un sentido geométrico, los dos puntos p_i y p_j son vecinos de Gabriel, si el circulo con diámetro igual a la distancia entre p_i y p_j no contiene ningún otro punto $p_k \in V$.

La técnica Grafo de Vecindad Relativa (*RNG*) [18], basa su funcionamiento en que si dos punto *x, y* son relativamente cercanos sí el área definida por la intersección del circulo con centro *x* y radio *xy*, y el circulo con centro en *y* con el mismo radio no contiene otro punto. Formalmente, el grado de la vecindad relativa de un conjunto de puntos *S* tiene una arista entre *x* e *y* si:

$$\text{dist}(x, y) \leq \max[\text{dist}(x, z), \text{dist}(y, z)] \quad \forall z \in S, \; z \neq x, y \qquad (\mathbf{4})$$

La desventaja de estos métodos es que no se tiene control sobre los patrones que se eliminan de la ME, y su uso en algunos casos aumenta el índice de error en proceso de aprendizaje [2].

4 Conclusiones

En este artículo concluimos que la complejidad de los datos son todos aquellos factores que afectan la calidad de la muestra de entrenamiento y por consiguiente disminuyen la presión de clasificación de la RNA, estos factores son el ruido, patrones atípicos, solapamiento de clases, desbalance de clases, tamaño de la ME y la distribución de los datos.

Hemos discutido sobre el problema llamado complejidad de los datos y sus efectos en el entrenamiento de una red neuronal. También, describimos las principales soluciones utilizadas para tratar este problema, lo que nos ayuda a comprender la magnitud de este problema, así como la dirección que deben tomar las investigaciones futuras respecto a este.

Algunas investigaciones futuras que podemos abordar como parte de este artículo es la búsqueda de métricas que nos ayuden a medir los factores englobados en la complejidad de los datos, así como la modificación de las técnicas anteriores para obtener mejores resultados, como por ejemplo, hasta ahora varios de estos métodos mencionados se han aplicado en el espacio de entrada, pero ¿qué pasaría si se aplicaran en el espacio oculto de la red neuronal?, que es donde en realidad trabaja la red, ¿se mejoraría el desempeño de la red?.

Bibliografía

1. Foody, G. M., *Using prior knowledge in artificial neural network with a minimal training set*, Int. Journal of Remote Sensing, 16, 2, 301-312, (1995).

2. Visa, S., Ralescu, A., *Issues in Mining Imbalanced Data Sets - A Review Paper*, Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, 67-73, (2005).

3. Barandela, R., Gasca, E., Alejo R., *Corrección de la muestra para el aprendizaje del Perceptron Multicapa*, Revista Iberoamericana de Inteligencia Artificial No. 13, 2{9, (2001).

4. Barandela, R., Gasca, E., Alejo R., *Correcting the training data. Combinatorial Optimization*, -Dordrecht- 13, 1-4-2,(2002).

5. Anand, R., Mehrotra, K., Mohan, C., Ranka, S., *An improved algorithm for neural network classification of imbalanced training sets*, IEEE Transactions on Neural Networks, 4, (1993).

6. Murphey, Y., Guo, H., Feldkamp, L., *Neural learning from imbalanced data*, Applied Intelligence, 21, (2004).

7. Bruzzone, L., Serpico, S., *Classification of imbalanced remote-sensing data by neural networks*, Pattern Recognition Letter, 18, (1997).

8. Lu, Y., Guo, H., Feldkamp, L., *Robust neural learning from unbalanced data examples*, In: Proc. of IEEE International Joint Conference on Neural networks, (1998).

9. Sanchez, J. S., Mollineda, R. A., Sotoca, J. M., *An analisis of how training data complexity affects the nearest neighbor classifiers*, Springer-Verlag London, (2007).

10. Haibo He, Garcia, Edwardo A., *Learning from Imbalanced Data*, IEEE Transactions on knowledge and Data Engineering, Vol. 21, No. 9, (2009).

11. Zhi-Hua, Xu-Ying Liu, *Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem*, IEEE Transactions on knowledge and Data Engineering, Vol. 18, No. 1, (2006).

12. Sánchez, Garreta, José Salvador, *Training Aprendizaje y clasificación basados en criterios de vecindad: Métodos alternativos y análisis comparativo*, Tesis Doctoral, Universitat Jaume I, (1998).

13. D. Mease, A. J., Wyner, and A. Buja, *Boosted Classification Trees and Class Probability/Quantile Estimation*, J. Machine Learning Research, Vol. 8, (2007).

14. Kukar, M., Kononenko, I., *Cost-sensitive learning with neural networks*, In: 13th European Conference on Artificial Intelligence, (1998).

15. Gopalakrishnan, M., V. Sridhar y H. KrishNamurthy. *Some application of clustering in the desing of neural networks*, Patter Recognition Letters, 16, 59-65, 1995.

16. Barandela, R y E. Gasca. *Decontamination of training samples for supervised pattern recognition methods*. En: Advances in Pattern Recognition, F. Ferri et al. (eds), Lecture Notes in Computer Science, 1876, Springer, 2000.

17. Wilson, D.L.: *Asymptotic properties of nearest neighbor rules using edited data sets*. IEEE Trans. on Systems, Man and Cybernetics 2 408-421, 1972.

18. Sánchez,J.S., F,Pla and F.J.Ferri, *Using the nearest centroid neighbourhood concept for editing purpose*, In Proc. VII Simposium Nacional de Reconocimiento de formas y Análisis de Imágenes 1, 175-180, 1997.

18. T. G. Dietterich, *Ensamble Learning*, The Handbook of Brain Theory and Neural Networks, second ed., M.A. Arbib, ed., Cambridge, Mass: MIT Press, 2002.