KDD CUP 2004 Approach, Quantum Physics Problem

Sergio Jiménez Vargas

Universidad Nacional de Colombia
Cod : 299655

1. INTRODUCCION

El conjunto de datos *Quantum Physics Problem* de la *KDD CUP 2004,* es un problema de clasificación con 78 atributos.y una clase binaria. Todos los atributos son numéricos y todos los valores se encuentran en general en el intervalo (-10,10). Existen 8 atributos que tienen cantidades considerables de valores faltantes. El conjunto de entrenamiento tiene 50.000 ejemplos y el conjunto de prueba tiene 100.000 ejemplos.

El enfoque para obtener un conjunto de prediciciones para el conjunto de prueba, fue probar diferentes clasificadores de la herramienta Weka [WEKA]. Se probaron técnicas de preprocesamiento, selección de atributos, clasificación, clustering y metaclasificadores.

En la anterior entrega parcial de este trabajo el mejor resultado obtuvo el puesto 44

2. ULTIMO MEJOR RESULTADO OBTENIDO

En la anterior entrega parcial de este trabajo el mejor resultado obtuvo el puesto **44** en *Accuracy* (Figura 1) y un promedio de puesto para las otras medidas de rendimiento de **59.5**. El enfoque utilizado para ese resultado fue un clasificador de regresión ***Logistics*** combinado con una pre-selección de atributos ***Chi-Squared.***

7	-	40.0	0.71096	37.0	0.79646
869	9	41.0	0.71094	56.0	0.75944
585	-	42.0	0.70972	35.0	0.79748
27	1	[illegible]	0.70961	34.0	0.79779
Sergio Jimenez	11	44.0	0.70943	65.0	0.70937
870	1	[illegible]	0.70901	-	-
711	1	46.0	0.70885	36.0	0.79731
318	-	47.0	0.70858	-	-

Figura 1. Mejor resultado *Accuracy* KDD CUP 2004 (entrega parcial octubre 2006)

En la entrega anterior, también se concluyó que los mejores clasificadores para el conjunto de datos utilizando *Cross-validation 10-folds* fueron *Logistics, J48* (árbol) y *Multilayer-Perceptron.* Los resultados finales para las medidas de rendimiento *Accuracy* y *ROC-area* se muestran en la Figura 2 y en la Tabla 1.

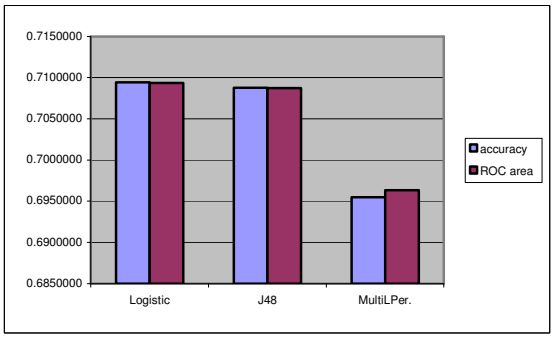

Figura 2. Mejores clasificadores KDD CUP 2004

	Accuracy	***ROC area***
Logistic	0.70943	0.70937
J48	0.70877	0.70874
MultiLPerceptron	0.69550	0.69633

Tabla 1. Mejores clasificadores KDD CUP 2004 (entrega parcial octubre 2006)

Las medidas de rendimiento *Cross-Entropy* y *SLQ*, se obtuvieron resultados no aceptables con los *ranking*s en el final de la lista.

3. PREPROCESAMIENTO NO SUPERVISADO

Se hicieron las siguientes pruebas de preprocesamiento no supervisado: *binarization*, *normalize,* y *replace-missing-values.* Entre las anteriores solamente *replace-missing-values* no empeoró la precisión de los clasificadores *Logistic, J48* y *MultiLayerPerceptron,* manteniendo el mismo resultado para las pruebas con el conjunto de entrenamiento completo y con *CrossValidation-10Folds.*

Dado que la precisión de los clasificadores se mantuvo y que el reemplazo de valores faltantes es conveniente para muchos clasificadores, se decidió trabajar en adelante con todos los valores faltantes reemplazados por la media de su respectivo atributo.

4. PREPROCESAMIENTO SUPERVISADO Y SELECCION DE ATRIBUTOS

En la entrega parcial anterior se encontró que la selección de atributos *Chi-Squared* obtuvo buenos resultados. Esta selección de atributos consiste en calcular la estadística Chi-cuadrado de cada atributo versus la clase y así obtener del nivel de correlación entre la clase y cada atributo. Dados los buenos resultados

obtenidos se pretende continuar con el uso de dicha selección de atributos.

Como preprocesamiento supervisado, se probó la discretización supervisada de *Fayyad & Irani* (2) que encuentra un conjunto de *bins* buscando maximizar la medida de disparidad *InfoGain* entre cada atributo y la clase. Dado que este método de discretización puede obtener un solo *bin* para un atributo dado, esto se puede considerar como una selección de atributos descartando aquellos a los que se encontró un solo *bin.*

En la Tabla 2 se compara la selección de atributos obtenida con *Chi-Squared* y con la discretización de *Fayyad & Irani.*

Atributo	Número de bins discretización de Fayyad & Irani	Estadística Chi-cuadrado atributo vs. Clase
F4	3	843.802
F8	3	1119.816
F12	2	530.232
F13	10	8307.477
F14	9	684.363
F15	5	505.170
F20	9	5122.202
F31	3	1874.660
F56	3	1679.978
F63	15	5561.208
F66	3	3087.818
F69	3	0.000
F70	7	4595.842
F71	10	3525.986
F75	3	66.628
otros atributos	1	0.000

Tabla 2. Comparación discretización de *Fayyad & Irani* vs. selección de atributos con la estadística *Chi-Squared.*

Dado que solamente hubo discrepancia entre los dos métodos en cuanto al atributo *F69*, se decidió incluir ese atributo en adelante a pesar que el método *Chi-squared* lo descarta.

Los intervalos obtenidos en la discretización tuvieron que ser aplicados al conjunto de prueba programando una rutina, ya que la herramienta *WEKA* no puede preprocesar los datos ya que es un método supervisado y los datos de prueba no tienen el atributo *class*.

Debido a un error en la asignación de los datos a los intervalos en la herramienta *WEKA* se tuvo que descartar la asignación de intervalos en el conjunto de entrenamiento. En su lugar se utilizó la misma rutina con la que se asignaron los intervalos en el conjunto de prueba.

La discretización supervisada de *WEKA* falla al asignar comparar los datos con los intervalos aparentemente por un error de precisión. Un ejemplo del error se tiene en el ejemplo #24 del conjunto de entrenamiento para el atributo *F14*. El valor del atributo es *0.000001* el cual es asignado al intervalo *(0.000001,0.000057]* cuando realmente pertenece al intervalo *(-0.000001, 0.000001]*.

5. PRUEBA DE CLASIFICADORES

La evaluación del rendimiento de los clasificadores se realizó considerando únicamente la medida de precisión. Los clasificadores a probar fueron: *Logistic, J48* y *MultilayerPerceptron.* Para el árbol *J48* se determinó un factor de poda óptimo de *0.218* utilizando el conjunto de datos de entrenamiento. Para *MultilayerPercepton* se determinó de igual forma un *learning-rate* de *0.2* y se probaron con *10* y *20* iteraciones.

En la Tabla 3 se muestran los resultados obtenidos con el conjunto de entrenamiento, con validación cruzada *10-folds* en el conjunto de entrenamiento y la medida de *Accuracy* obtenida en el sitio de la *KDD CUP 2004* para las predicciones.

	Accuracy Training-Set	Crossvalidation 10-Folds	KDD CUP Accuracy
Logistic	72.538%	72.290%	72.151%
J48	74.308%	71.734%	71.489%
MultiLPer. (10 iteraciones)	71.918%	71.156%	71.034%
MultiLPer. (20 iteraciones)	72.712%	no se calculó	71.012%

Tabla 3. Resultados de *Accuracy* para los mejores clasificadores.

Todos los clasificadores (*Logistic, J48*, y *MultiLayerPerceptron*) superaron el mejor resultado obtenido en la entrega parcial anterior de este proyecto, donde la mejor precisión fue *70.943%* para la medida *Accuracy* de la *KDD CUP 2004.*.

6. CLUSTERING

Se evaluaron diferentes algoritmos de *clustering* y conjuntos de datos. Los datos utilizados fue el conjunto de entrenamiento con reemplazo de los valores faltantes y se hicieron pruebas con todos los atributos y los seleccionados con el criterio *Chi-squared* mostrados en la Tabla 2.

La discretización de *Fayyad & Irani*, no se utilizó ya que redujo la precisión de predicción de la *class* con los *clusters* en todos los casos. Este resultado era predecible ya que el proceso de discretización introduce ruido en el cálculo de las distancias entre ejemplos.

Los resultados de las pruebas de *clusteríng* se muestran en la Tabla 4.

Conjunto de datos	Algoritmo	Precisión vs. Class
Conjunto de prueba, missing-replace, Fayyad & Irani discretization	Simple K-means	56.44%
Conjunto de prueba, missing-replace, Chi-squared	Simple K-means	61.65%
Conjunto de prueba, missing-replace	Simple K-means	50.14%
Conjunto de prueba, missing-replace, Chi-squared	Make Density Based Cluster (K-means)	61.92%

Conjunto de prueba, missing-replace, Chi-squared	Farthest First	54.61%

Tabla 4. Resultados de *clustering* para el conjunto de datos de prueba en comparación a la clase

El mejor resultado se obtuvo con el algoritmo *Make Density Based Cluster* con *K-means.* Los *clusters* encontrados se adicionaron como un atributo más para los datos, pero en ninguno de los clasificadores que se probaron en la Tabla 3 se obtuvo mejores resultados.

7. METACLASIFICADORES

Los metaclasificadores que se probaron fueron: *Bagging, AdaBostM1* y *.MultiBoostAB.* El único que logró mejorar la medida de precisión de clasificación fue combinando *Bagging* y el clasificador *Logistic.* Los resultados mejoraron levemente las medidas de rendimiento del clasificador *Logistic* y se muestran en la Tabla 5.

	Accuracy (KDD CUP)	***ROC area (KDD CUP)***
Logistic	72.151%	0.80929
Bagging con Logistic	72.170%	0.81269

Tabla 5. Resultados de la utilización de *Bagging* para el clasificador *Logistic*

8. AJUSTES PARA *CROSS-ENTROPY* Y *SLQ*

Todos las prediciones que se obtuvieron hasta este punto, generaron prediciones para la clase en el conjunto *{0, 1}*. Sin embargo para el calculo de la medida de rendimiento *CrossEntropy*, la cual se define en (1) donde *tar* es el valor desconocido para nosotros de la clase y *pred* es la predicción que se hace. En esta definición se requiere una probabilidad en el intervalo abierto (0,1).

$$CrossEntropy = \sum [tar \log(pred) + (1 - tar) \log(1 - pred)]$$

WEKA provee dicha probabilidad para el clasificador *Logistic, J48* y *Multilayer-Perceptron.* Sin embargo se obtienen valores para 0 y 1. El valor de *0* se reemplazó por *0.0001* y el valor de *1* se reemplazó por *0.9999.*

9. RESULTADO FINAL

El mejor resultado se obtuvo aplicando las siguientes técnicas al conjunto de entrenamiento y usando el modelo obtenido para predecir en el conjunto de prueba:

- Reemplazo de valores faltantes por la media
- Discretización de *Fayyad & Irani*
- Clasificador de regresión *Logistic*
- MetaClasificador *Bagging*

La Tabla 6 muestra los resultados y el ranking obtenido para cada métrica. La Figura 2 muestra una copia de la página web de la conferencia *KDD CUP 2004,* con los 13 mejores resultados.

Métrica	***Valor***	***Ranking***
Accuracy	0.72170	14
ROC area	0.81269	13
Cross Entropy	0.74339	12
SLQ	0.29900	13
Average Rank		13
Mejor solución		**11**

Tabla 6. Mejor resultado obtenido

Group	New Subs.	Accuracy		ROCA		Cross Entropy		SLQ	Score	Avg Rank
MEDai/AI Insight	2	2.0	0.73120	1.0	0.82774	1.0	0.71426	1.0	0.32775	1.250
Inductis	-	1.0	0.73255	2.0	0.82754	2.0	0.72456	2.0	0.32648	1.750
Golden Helix	-	4.0	0.72775	3.0	0.82250	6.0	0.73001	4.0	0.31749	4.250
Andre and Tiny	55	3.0	0.72799	5.0	0.82158	5.0	0.72853	5.0	0.31528	4.500
Mario Ziller '	1	9.0	0.72451	4.0	0.82243	3.0	0.72516	3.0	0.31758	4.750
Ahmad abdulKader	-	6.0	0.72744	9.0	0.81791	4.0	0.72798	8.0	0.30982	6.750
Salford Systems	-	5.0	0.72745	6.0	0.82109	13.0	0.74371	6.0	0.31447	7.500
Probing - JL&BZ	-	7.0	0.72522	8.0	0.81906	8.0	0.73634	7.0	0.31142	7.500
FEG, Japan	-	15.0	0.72060	11.0	0.81572	7.0	0.73583	10.0	0.30591	10.750
Inductis_Team2	1	10.0	0.72398	12.0	0.81442	11.0	0.74101	12.0	0.30197	11.250
Sergio Jimenez	33	14.0	0.72170	13.0	0.81269	12.0	0.74339	13.0	0.29900	13.000
[illegible]	-	12.0	0.72304	14.0	0.81252	15.0	0.74632	11.0	0.30410	13.000
Tiberius	1	8.0	0.72509	16.0	0.81116	19.0	0.75301	14.0	0.29569	14.250

Figura 2. Ranking KDD CUP 2004

REFERENCIAS

(1) WEKA. University of Waikato
http://www.cs.waikato.ac.nz/ml/weka

(2) Usama M. Fayyad and Keki B. Irani, *On the Handling of Continuous-Valued Attributes in Decision Tree Generation*.Machine Learning 1992 vol 8 pags. 87102

(3) KDD CUP 2004. *KDD Conference 2004*
http://kodiak.cs.cornell.edu/kddcup