

Conjuntos con ruido

Se realizó un análisis comparativo sobre 5 métodos de manejo de ruido sobre 3 bases de datos con valores 4 niveles de ruido, 5 %, 10 %, 15 % y 20 %, donde la forma de medir la efectividad del método de imputación fue realizar una tarea de clasificación con 3 métodos de clasificación diferentes, árboles de decisión, vecinos más cercanos (KNC), perceptrón multicapa (NN) y máquinas de soporte vectorial (SVC), y para comparar los métodos se hizo uso del estadístico Nemenyi Friedman, obteniendo sus valores AUC, donde una mayor área nos sugiere un mejor modelo, los resultados obtenidos se muestran en los cuadros 2,3,4 y 5.

Para que este análisis fuese robusto se realizaron múltiples réplicas de validación cruzada, 5 2 veces validación cruzada, usando el método stratified cross validation de Sklearn, estos subconjuntos sirvieron para entrenar los tres modelos de clasificación, para una mejor comparación se obtuvieron los ranks de cada observación de los cuadros 2-5 y para una comparación más sencilla se obtuvieron los promedios que se muestran en el cuadro 1.

El cuadro 1 muestra que para las bases con 5 % de ruido el mejor método de filtración es el EF, seguido del ENN, mientras que para el resto los dos mejores son ENN e IEKNCN.

	ENN	EF	IEKNCN	DROP	CVCF
5 %	2.72	2.39	3.11	3.5	3.28
10 %	2.17	3.44	2.56	3.28	3.56
15 %	2.28	2.61	2.78	3.56	3.78
20 %	2.33	3.11	2.89	3.83	2.83

Cuadro 1: Rankings promedio para comparación.

Classifier/Method	ENN	EF	IEKNCN	DROP	CVCF	Database
Decision Tree	0.861	0.649	0.725	0.865	0.865	Heart
Decision Tree	0.920	0.920	0.920	0.920	0.920	Heart
Decision Tree	0.836	0.885	0.917	0.885	0.762	Heart
SVC	0.563	0.491	0.365	0.563	0.435	Pima
SVC	0.719	0.361	0.720	0.522	0.723	Pima
SVC	0.765	0.323	0.684	0.750	0.759	Pima
KNC	0.291	0.471	0.532	0.419	0.508	Ring
KNC	0.752	0.816	0.865	0.877	0.659	Ring
KNC	0.328	0.417	0.407	0.443	0.444	Ring
Promedio	0.671	0.593	0.682	0.694	0.675	

Cuadro 2: AUC para las bases con 5 % de ruido.

Classifier/Method	ENN	EF	IEKNCN	DROP	CVCF	Database
Decision Tree	0.465	0.585	0.445	0.577	0.454	Heart
Decision Tree	0.885	0.920	0.846	0.885	0.917	Heart
Decision Tree	0.785	0.608	0.557	0.814	0.817	Heart
SVC	0.412	0.482	0.423	0.496	0.480	Pima
SVC	0.442	0.827	0.776	0.795	0.802	Pima
SVC	0.741	0.624	0.806	0.524	0.798	Pima
KNC	0.322	0.423	0.507	0.387	0.544	Ring
KNC	0.638	0.614	0.784	0.618	0.507	Ring
KNC	0.328	0.415	0.412	0.443	0.437	Ring
Promedio	0.558	0.611	0.617	0.615	0.639	

Cuadro 3: AUC para las bases con 10 % de ruido.

Classifier/Method	ENN	EF	IEKNCN	DROP	CVCF	Database
Decision Tree	0.271	0.583	0.583	0.371	0.615	Heart
Decision Tree	0.517	0.795	0.827	0.844	0.810	Heart
Decision Tree	0.328	0.421	0.403	0.443	0.453	Heart
SVC	0.556	0.698	0.486	0.598	0.486	Pima
SVC	0.865	0.649	0.861	0.725	0.865	Pima
SVC	0.473	0.500	0.599	0.556	0.699	Pima
KNC	0.609	0.593	0.467	0.538	0.369	Ring
KNC	0.760	0.684	0.758	0.777	0.825	Ring
KNC	0.540	0.629	0.635	0.735	0.719	Ring
Promedio	0.547	0.617	0.624	0.621	0.649	

Cuadro 4: AUC para las bases con 15 % de ruido.

Classifier/Method	ENN	EF	IEKNCN	DROP	CVCF	Database
Decision Tree	0.222	0.476	0.678	0.587	0.674	Heart
Decision Tree	0.400	0.817	0.782	0.782	0.724	Heart
Decision Tree	0.328	0.438	0.383	0.445	0.469	Heart
SVC	0.562	0.563	0.390	0.566	0.381	Pima
SVC	0.833	0.751	0.735	0.650	0.750	Pima
SVC	0.757	0.566	0.693	0.783	0.767	Pima
KNC	0.662	0.629	0.667	0.791	0.495	Ring
KNC	0.920	0.920	0.920	0.920	0.920	Ring
KNC	0.418	0.679	0.646	0.763	0.646	Ring
Promedio	0.567	0.649	0.655	0.699	0.647	

Cuadro 5: AUC para las bases con 20 % de ruido.