Propuesta de Mejora

Problema

Al incrementarse la cantidad de objetos clasificados en la base de datos se degrada la performance del algoritmo de clasificacion. A medida que se van haciendo clasificaciones y se guarda el resultado en la base de datos, se empieza a notar que la barra de progreso tarda cada vez mas en aparecer e iniciar la clasificacion.

Causa

Cada vez que se inicia una nueva clasificacion se calculan el valor medio y el desvio estandar para cada rasgo de una clase. Esto se hace consultando los valores del rasgo en todas los objetos de una clase dada, mediante los siguientes querys:

media:
```
select sum(r.valor) / count(*)
       from rasgo_objeto r
       join objeto o on o.uid = r.id_objeto
       join clase_objeto c on o.uid = c.id_objeto
       where c.id_clase = ID_CLASE
       and r.id_rasgo = ID_RASGO;
```

desvio estandar:
```
select sqrt(sum(power(r.valor,2) ) / count(*) -
          power(sum(r.valor) / count(*) ,2))
      from rasgo_objeto r
      join objeto o on o.uid = r.id_objeto
      join clase_objeto c on o.uid = c.id_objeto
      where c.id_clase = ID_CLASE
      and r.id_rasgo = ID_RASGO;
```

A medida que se insertan mas objetos clasificados estos querys se van haciendo mas lentos debido a que la cantidad de objetos se incrementa.

Mejora 1

Una mejora es no calcular el valor medio y desvio de los rasgos de una clase cada vez que se inicia una clasificacion, sino calcularlo cuando se guarda la clasificacion en la base de datos, que es el momento donde se insertan los nuevos objetos. Con esta mejora no se degrada el tiempo de clasificacion pero mueve el problema a la operacion de guardado de los objetos en la base de datos, ya que despues guardar los objetos se deben calcular el valor medio y desvio estandar de los rasgos de una clase, y a medida que se incrementa la cantidad objetos se incrementa el tiempo de calculo.

Mejora 2

Para evitar que a medida que se agreguen mas datos el calculo del valor medio y desvio estandar tarden cada vez mas se puede calcular en forma incremental utilizando estas formulas :

$$media = \sum valor / N \qquad \text{y} \qquad desv_{est} = \sum valor^2 - media^2$$

Deberiamos guardar la sumatoria de los valores del rasgo para los objetos de una clase (sum_valor), la sumatoria de los valores al cuadrado de los objetos de una clase (sum_valor_cuadrado) y la cantidad de valores (N). Luego calcular el valor madio y desvio estandar usando esos valores.

Al momento de guardar los objetos de una clase se debe calcular los campos sum_valor y sum_valor_cuadrado de los objetos nuevos clasificados, y adicionarselos a los ya calculados previamente.

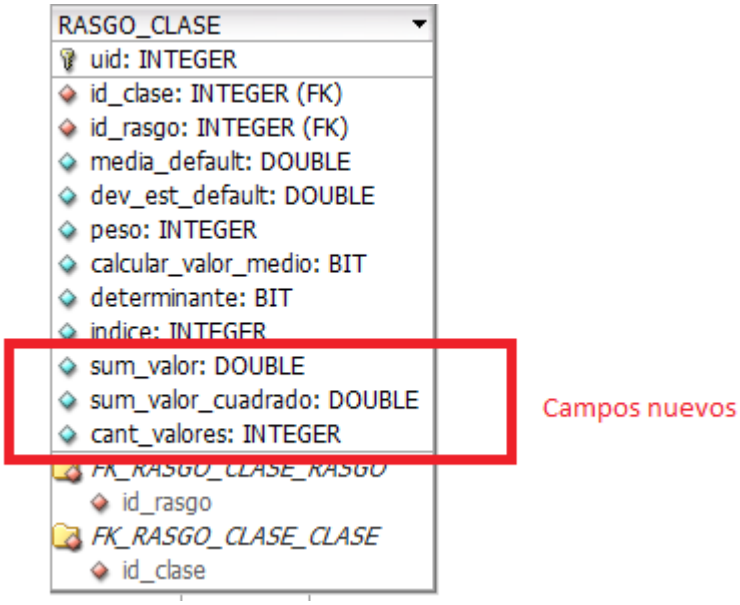
sum_valor = sum_valor + sum_valor_rasgo_nuevos

sum_valor_cuadrado = sum_valor_cuadrado + sum_valor_cuadrado_nuevos

Luego usar las formulas de valor medio y desvio estandar de arriba y actualizar esos valores en la base de datos.

Solucion final

Combinar las mejoras 1 y 2 : Calcular el valor medio y desvio estandar de los rasgos de una clase al momento de guardar la clasificacion y hacerlo en forma incremental.



Modificar la tabla rasgo_clase para agregar las columnas sum_valor, sum_valor_cuadrado y cant_valores. Luego el calculo de la media y desvio estandar se debe hacer en base a esas columnas.

Media = sum_valor / cant_valores;
desv_est = sum_valor_cuadrado / cant_valores – media ^ 2;

Luego al momento de guardar la clasificacion se actualizan los campos sum_valor, sum_valor_cuadrado y cant_valores de un rasgo de una clase.