**Bài toán**: Phân lớp đa nhãn (vẫn như bài toán cũ, nhưng sử dụng phương pháp khác)

Input: Dữ liệu đã gán nhãn, định dạng arff

Như vậy là qua file arff và xml mình sẽ phải xác định số nhãn và giá trị nhãn của mỗi dòng

Mình sẽ sử dụng file arff của mình và file arff của tác giả thuật toán cho để thực nghiệm và so sánh.

output: mô hình phân lớp 🡪 đánh giá mô hình trên tập test

**Các bước thực hiện:**

Từ tập arff và xml 🡪 xác định nhãn và giá trị nhãn của mỗi thể hiện nhé (mỗi thể hiện là 1 dòng)

Ví dụ: file 5.arff và 5.xml 🡪 có 5 nhãn được định nghĩa trong file xml.

và tại mỗi dòng, ví dụ dòng số 1:

{1 1,5 1,15 1,44 1,67 1,92 2,95 1,135 1,167 1,184 1,195 1,205 1,207 1,523 1,531 1,756 1,785 1,992 1,1750 1,1751 1,1754 1}

cần xác định được dòng này có nhãn 1,2,5

// giống như dịch ngược í nhỉ

1. Đối với mỗi nhãn lk, xác định tập âm, dương trên dữ liệu. Trong đó:

+ Tập dương là các dòng có nhãn lk

+ Tập âm là các dòng ko có nhãn lk

Ví dụ: trong file arff nhãn lk là đặc trưng thứ m

thì trong phần @data: dòng nào có giá trị tại đặc trưng thứ m thì dòng đó xếp vào tập dương, dòng nào ko có giá trị tại đặc trưng thứ m thì nghĩa là ko có nhãn đó và xếp vào tập âm

Có bao nhiêu nhãn thì làm từng đấy lần bước này.

1. Với mỗi tập âm và dương của mỗi nhãn ở trên, tiến hành Phân cụm riêng rẽ trên mỗi tập em nhé.

Xác định số cụm theo công thức:

+ m = r . lực lượng tập âm . lực lượng tấp dương

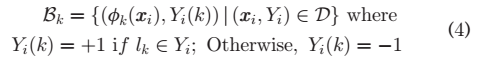
+ r là giá trị từ 0 🡪 1 (Cái này sẽ thử nhiều giá trị khác nhau)

+ Sử dụng phương pháp k-mean để phân cụm

1. Xây dựng ánh xạ bằng cách tính khoảng cách của các thể hiện với tâm các cụm



1. Với mỗi nhãn lk, xây dựng tập huấn luyện nhị phân với dạng sau:



Về bản chất là thêm giá trị 1 hoặc -1 vào cuối ánh xạ ở bước 3 tuỳ thuộc vào thể hiện đó có thuộc nhãn lk hay không :D

1. Áp dụng một máy phân lớp nhị phân, tác giả paper sử dụng CLR và ECC lên tập dữ liệu trên để thu được mô hình phân lớp
2. Tiến hành phân lớp trên tập dữ liệu test với mô hình trên theo công thức:



Trong đó, u là thể hiện cần phân nhãn trong tập test.

Kết quả phân lớp là tập các nhãn lk mà thoả mãn điệu kiện công thức 5

Dữ liệu format ARFF: Gồm 3 phần:  **@RELATION**, **@ATTRIBUTE** and **@DATA**

1. **@RELATION** định nghĩa tên của dữ liệu
2. **@ATTRIBUTE** định nghĩa các tham số, kiểu tham số và nhãn dữ liệu (thường nhãn ở dưới dùng)
3. **@DATA**: dữ liệu thực theo thứ tự tham số được định nghĩa ở @ATTRIBUTE

Ví dụ tập dữ liệu bitex:

1. @relation bibsonomy\_bibtex: tên của dữ liệu
2. @attribute: định nghĩa các tham số là các số và word theo định dạng {0,1} nghĩa là xuất hiện hay ko xuất hiện. Như vậy đặc trưng theo kiểu nhị phân

Phần cuối là định nghĩa nhãn: @attribute tag\_2005,… theo dạng {0,1} nghĩa là có hoặc ko có nhãn đó

1. @data là dữ liệu thực

Mỗi dòng là 1 văn bản được biểu diễn theo các attribute định nghĩa và nhãn mà tài liệu đó thuộc về.

Chú ý: Để xác định nhãn, dựa vào file xml của tệp arff đó em nhé.