Predizione della funzione delle proteine con metodi di Machine Learning

Marco Odore
Lorenzo Rossi

18 settembre 2017

Docente: Valentini Giorgio
Corso: Bioinformatica

1 Scopo del progetto

L'obiettivo del progetto è di predire la funzione delle proteine di *Drosophila melanogaster*, per determinate ontologie, tramite gli algoritmi di apprendimento Support Vector Machine(SVM) e Multilayer Perceptron(MLP), per poi analizzarne e confrontarne i risultati. Dato che ogni proteina può essere classificata in più di una categoria, il problema trattato è quello della classificazione multi-etichetta.

2 Dataset

Il dataset utilizzato per l'apprendimento induttivo è stato generato da un grafo indiretto, i cui nodi sono le proteine e gli archi indicano il grado di similitudine tra due proteine[1]. Tale grafo è rappresentato da una matrice pesata di adiacenza e ogni riga (colonna) si riferisce quindi ad una diversa proteina dell'organismo ed ogni entry al peso dell'arco che connette due proteine.

2.1 Istanze degli algoritmi

Le istanze utilizzate per i due algoritmi induttivi sono le righe (colonne) della matrice di adiacenza. Quindi per ogni proteina si avrà un vettore le cui componenti (feature) rappresentano il grado di similitudine che questa ha in relazione alle altre proteine.

[1] Come è stata costruita la matrice:
`https://homes.di.unimi.it/~valentini/SlideCorsi/Bioinformatica1617/Bioinf-Project1617.pdf`

2.2 Etichettatura

Per l'etichettatura delle istanze sono state fornite tre ontologie:

- BP(Biological Process) con 1951 termini.
- MF (Molecular Function) con 234 termini.
- CC (Cellular Component) con 235 termini.

Rappresentate da matrici di annotazioni, dove sulle righe si hanno le proteine e sulle colonne i termini delle ontologie.
Data la notevole quantità di tempo necessaria per l'addestramento dei classificatori, ci si è soffermati unicamente sull'ontologia CC. Quindi ad ogni istanza del problema è stato associato un sottoinsieme di queste etichette.

Stoichiometry The relationship between the relative quantities of substances taking part in a reaction or forming a compound, typically a ratio of whole integers.

Atomic mass The mass of an atom of a chemical element expressed in atomic mass units. It is approximately equivalent to the number of protons and neutrons in the atom (the mass number) or to the average number allowing for the relative abundances of different isotopes.

3 Experimental Data

Mass of empty crucible	7.28 g
Mass of crucible and magnesium before heating	8.59 g
Mass of crucible and magnesium oxide after heating	9.46 g
Balance used	#4
Magnesium from sample bottle	#1

4 Sample Calculation

Mass of magnesium metal	= 8.59 g - 7.28 g
	= 1.31 g
Mass of magnesium oxide	= 9.46 g - 7.28 g
	= 2.18 g
Mass of oxygen	= 2.18 g - 1.31 g
	= 0.87 g

Because of this reaction, the required ratio is the atomic weight of magnesium: 16.00 g of oxygen as experimental mass of Mg: experimental mass of oxygen or $\frac{x}{1.31} = \frac{16}{0.87}$ from which, $M_{\mathrm{Mg}} = 16.00 \times \frac{1.31}{0.87} = 24.1 = 24\,\mathrm{g\,mol^{-1}}$ (to two significant figures).

5 Results and Conclusions

The atomic weight of magnesium is concluded to be $24\,\mathrm{g\,mol^{-1}}$, as determined by the stoichiometry of its chemical combination with oxygen. This result is in agreement with the accepted value.

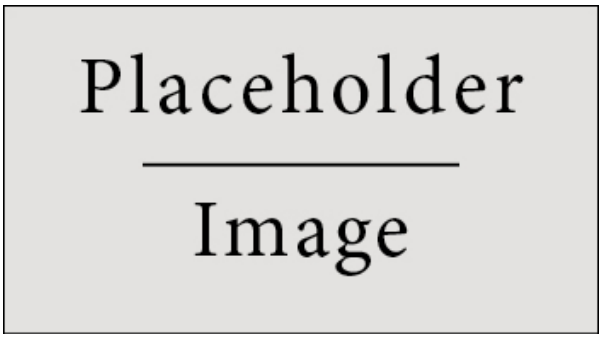

Figura 1: Figure caption.

6 Discussion of Experimental Uncertainty

The accepted value (periodic table) is $24.3\,\mathrm{g\,mol^{-1}}$ **?**. The percentage discrepancy between the accepted value and the result obtained here is 1.3%. Because only a single measurement was made, it is not possible to calculate an estimated standard deviation.
The most obvious source of experimental uncertainty is the limited precision of the balance. Other potential sources of experimental uncertainty are: the reaction might not be complete; if not enough time was allowed for total oxidation, less than complete oxidation of the magnesium might have, in part, reacted with nitrogen in the air (incorrect reaction); the magnesium oxide might have absorbed water from the air, and thus weigh "too much." Because the result obtained is close to the accepted value it is possible that some of these experimental uncertainties have fortuitously cancelled one another.

7 Answers to Definitions

a. The *atomic weight of an element* is the relative weight of one of its atoms compared to C-12 with a weight of 12.0000000..., hydrogen with a weight of 1.008, to oxygen with a weight of 16.00. Atomic weight is also the average weight of all the atoms of that element as they occur in nature.

b. The *units of atomic weight* are two-fold, with an identical numerical value. They are g/mole of atoms (or just g/mol) or amu/atom.

c. *Percentage discrepancy* between an accepted (literature) value and an experimental value is

$$\frac{\text{experimental result} - \text{accepted result}}{\text{accepted result}}$$