The Wisdom of Crowds
Cluster Ensemble

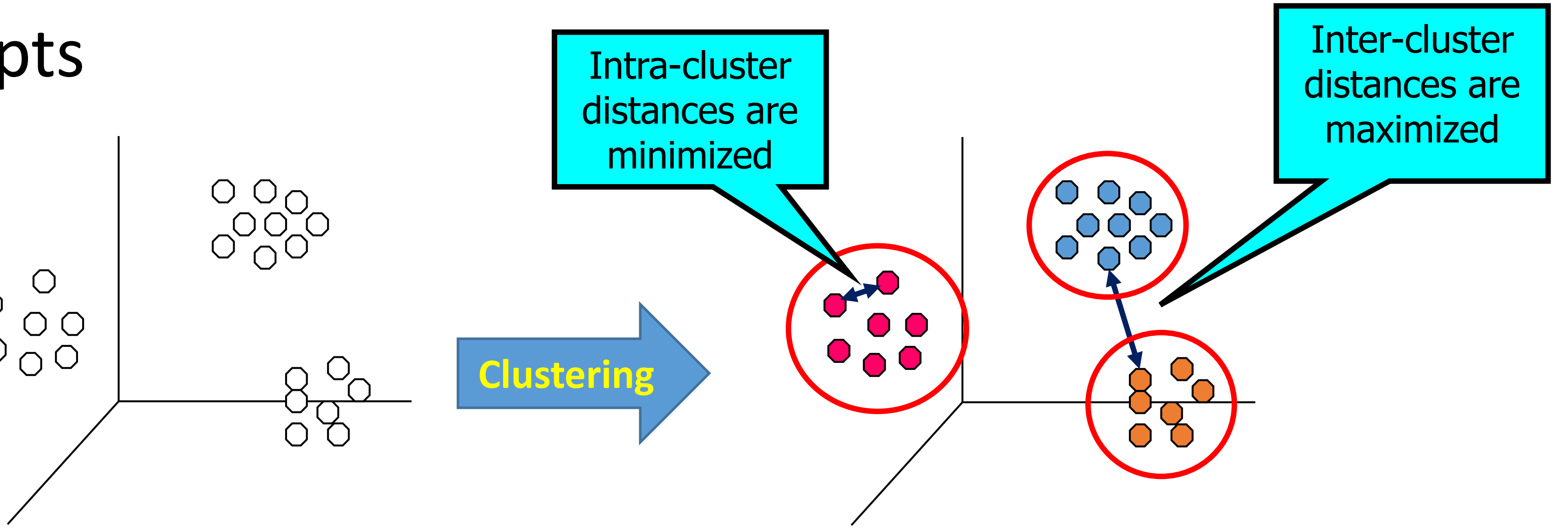
Muhammad Yousefnezhad
Member of ParNec & DQ groups
Department of Computer Science & Technology
Nanjing University of Aeronautics and Astronautics
email: myousefnezhad@nuaa.edu.cn

Outlines

- Concepts of Clustering
- Basic Clustering Algorithms
- Cluster Ensemble Selection
- Wisdom of Crowds Theory
- Wisdom of Crowds Cluster Ensemble (WOCCE)
- Experimental Results
- Feature Works

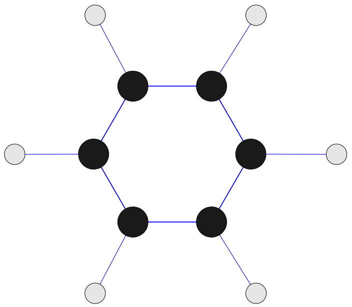
Concepts of Clustering

- Concepts

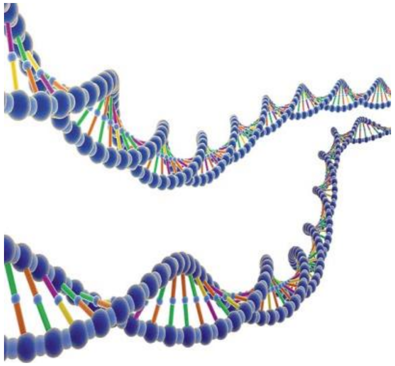


Intra-cluster distances are minimized
Clustering
Inter-cluster distances are maximized

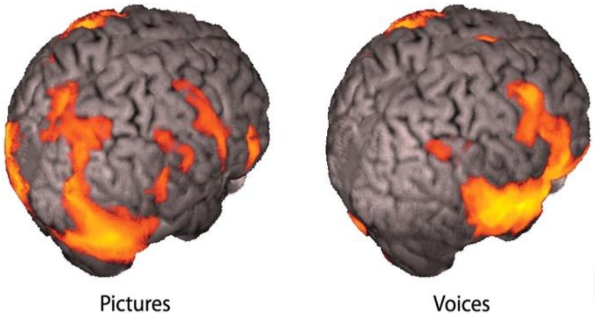
- Applications



Chemical Elements



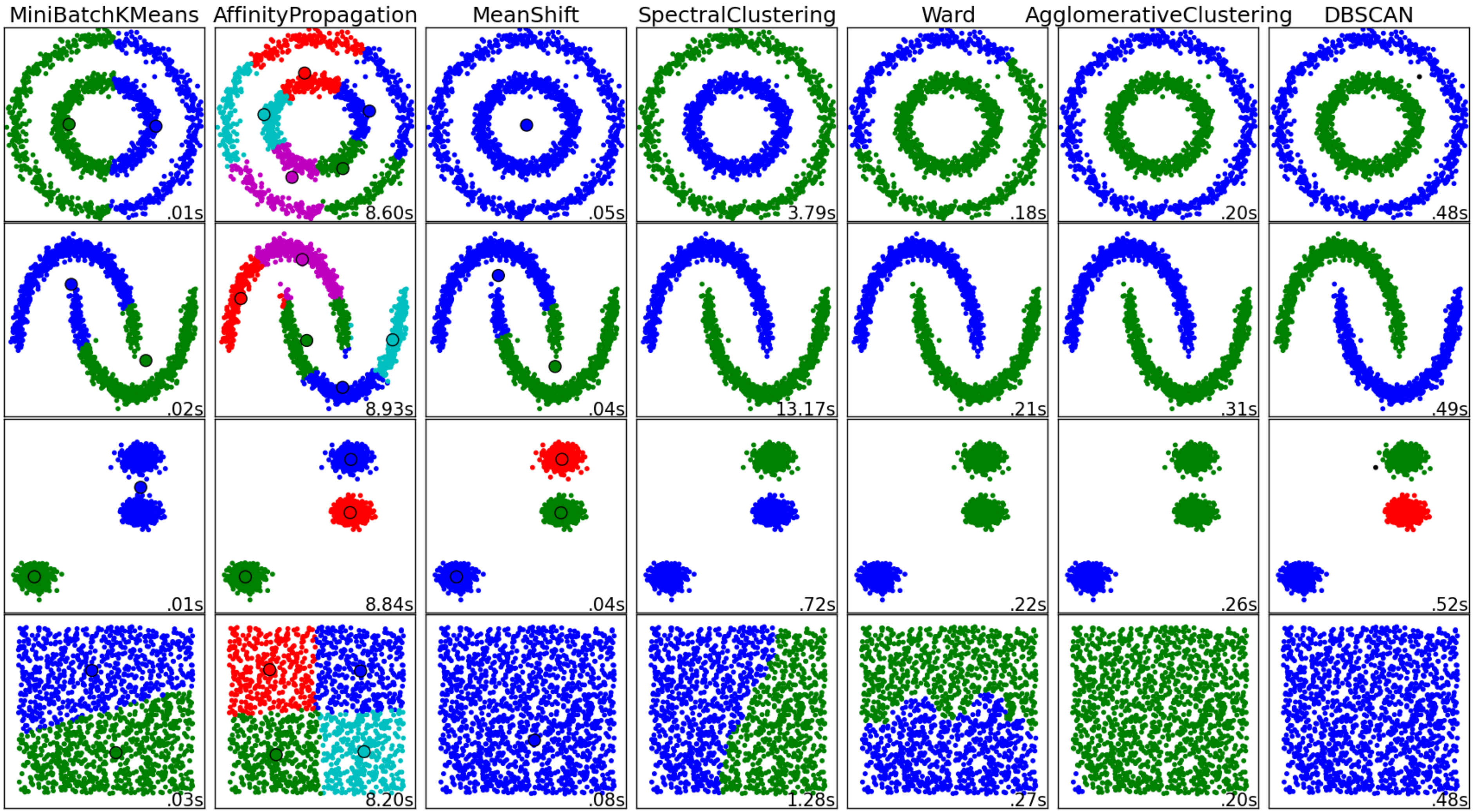
Gens



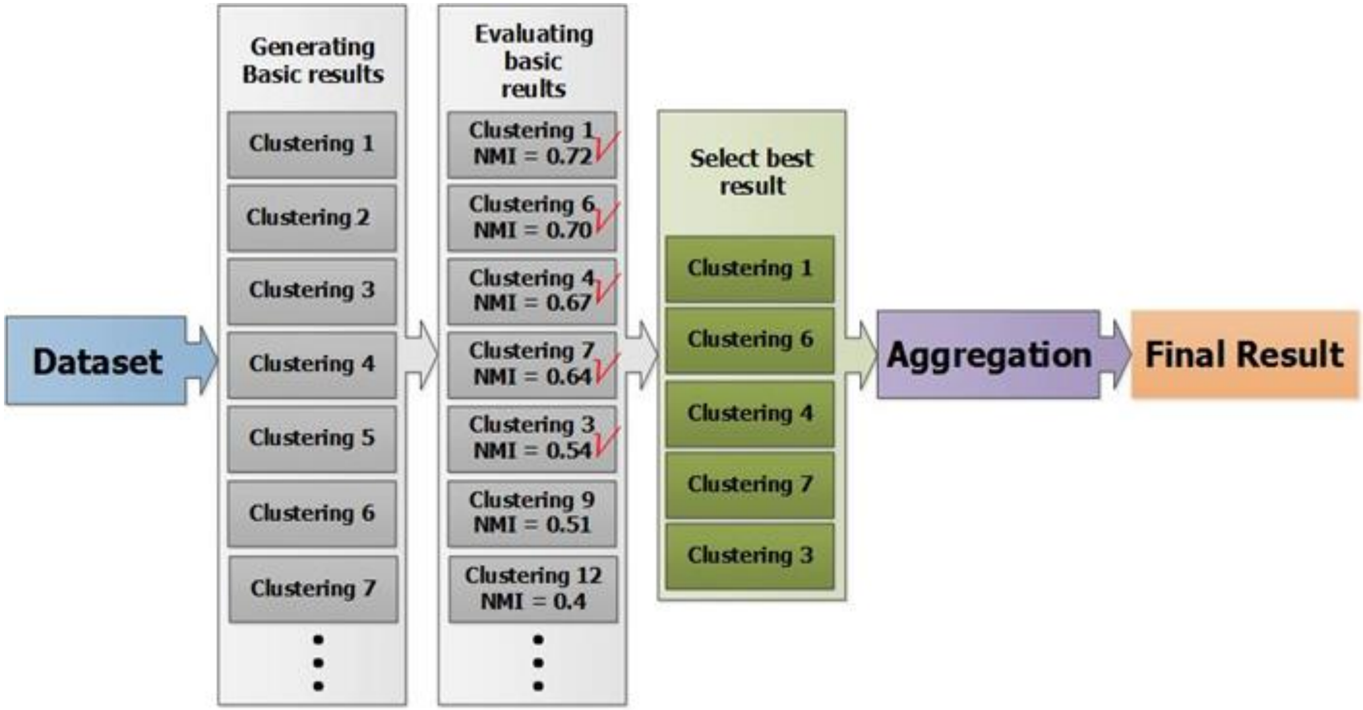
Brain Function Analysis

Web & Social Networks

Basic Clustering Algorithms



Cluster Ensemble Selection



Dataset

Generating Basic results

Clustering 1

Clustering 2

Clustering 3

Clustering 4

Clustering 5

Clustering 6

Clustering 7

⋮

Evaluating basic reults

Clustering 1 NMI = 0.72 ✓

Clustering 6 NMI = 0.70 ✓

Clustering 4 NMI = 0.67 ✓

Clustering 7 NMI = 0.64 ✓

Clustering 3 NMI = 0.54 ✓

Clustering 9 NMI = 0.51

Clustering 12 NMI = 0.4

⋮

Select best result

Clustering 1

Clustering 6

Clustering 4

Clustering 7

Clustering 3

Aggregation

Final Result

Challenges

- Strategy of Selection
- Generating Results
- Evaluation metrics
- Aggregation Methods

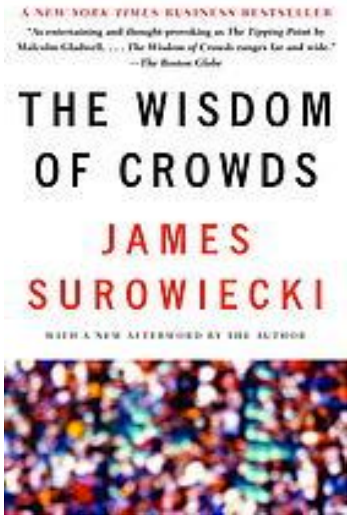
$$\text{Shannon } Entropy(X) = -\sum P(X)\log P(X)$$

Normal Mutual Information

$$NMI(P^a, P^b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab}\log\left(\frac{n_{ij}^{ab}\times n}{n_i^a \times n_j^b}\right)}{\sum_{i=1}^{k_a} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b} n_j^b \log\left(\frac{n_j^b}{n}\right)}$$

The Wisdom of Crowds (WOC) Theory

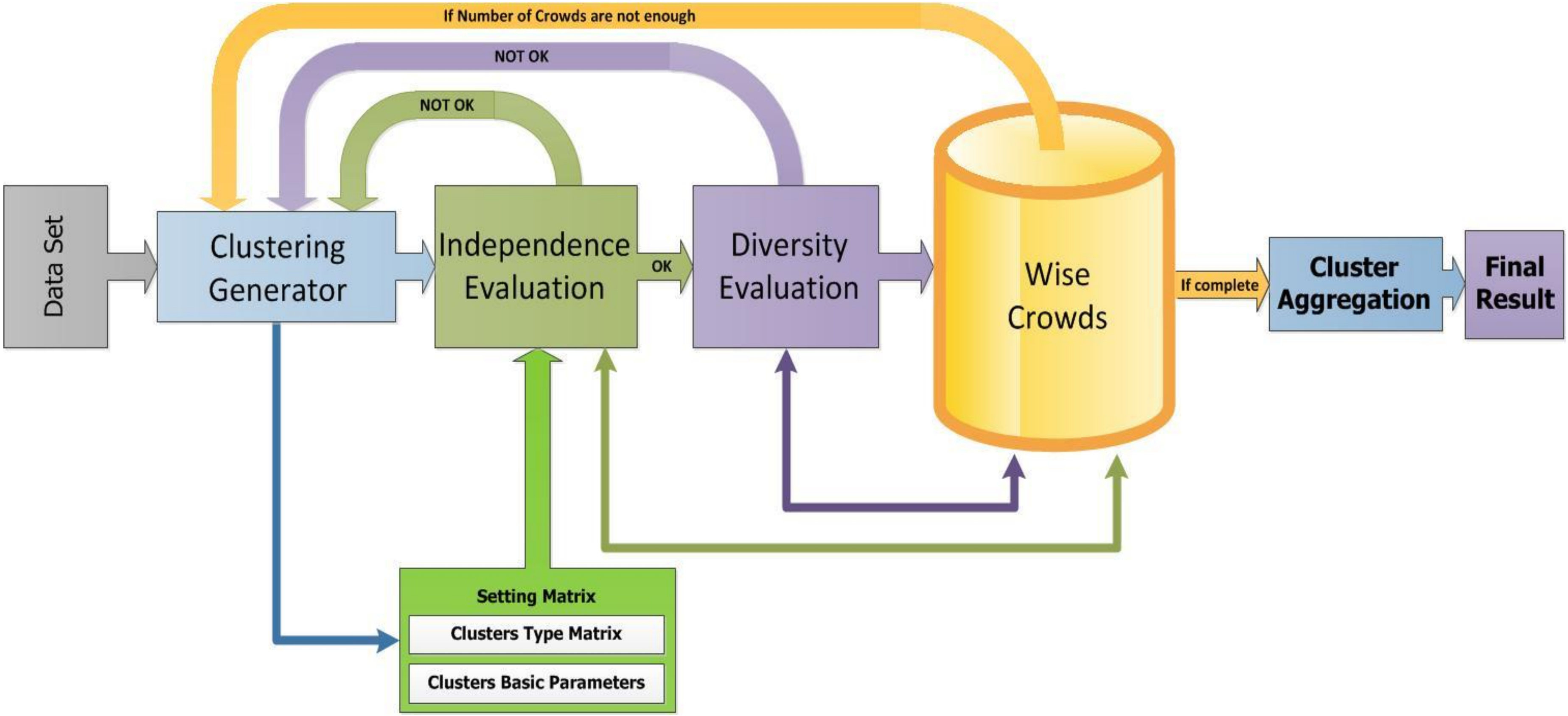
Surowiecki J. (2004), “The Wisdom of Crowds”, Cover of mass market edition by Anchor, *ISBN: 978-0385503860*.


A NEW YORK TIMES BUSINESS BESTSELLER
"As entertaining and thought-provoking as The Tipping Point by Malcolm Gladwell. . . . The Wisdom of Crowds ranges far and wide."
—The Boston Globe
THE WISDOM OF CROWDS
JAMES SUROWIECKI
WITH A NEW AFTERWORD BY THE AUTHOR

- **Background**
  - **Condorcet, (1785), “Jury Theorem”, Essay on the Application of Analysis to the Probability of Majority Decisions.**
  - **Galton, (1907), “Vox Populi”, Nature, vol. 75, pp. 450-451.**
- **Criteria**
  - **Diversity of opinion**
    - **Each person should have private information even if it's just an eccentric interpretation of the known facts.**
  - **Independence**
    - **People's opinions aren't determined by the opinions of those around them.**
  - **Decentralization**
    - **People are able to specialize and draw on local knowledge.**
  - **Aggregation**
    - **Some mechanism exists for turning private judgments into a collective decision.**
- **Application**
  - **Delphi Method, Crowd Sourcing, Machine Learning, and etc.**

Wisdom of Crowds Cluster Ensemble

The WOCCE framework (with feedback mechanism)



WOCCE – Diversity

$$APMM(C_i^a, P^{b*}) = \frac{-2\log\left(\frac{n}{n_i^a}\right)\sum_{j=1}^{k_{b*}} n_j^{b*}}{n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)}$$

$$AAPMM(P) = \frac{1}{M}\sum_{j=1}^{M} APMM\left(C, P_j^{b*}\right)$$

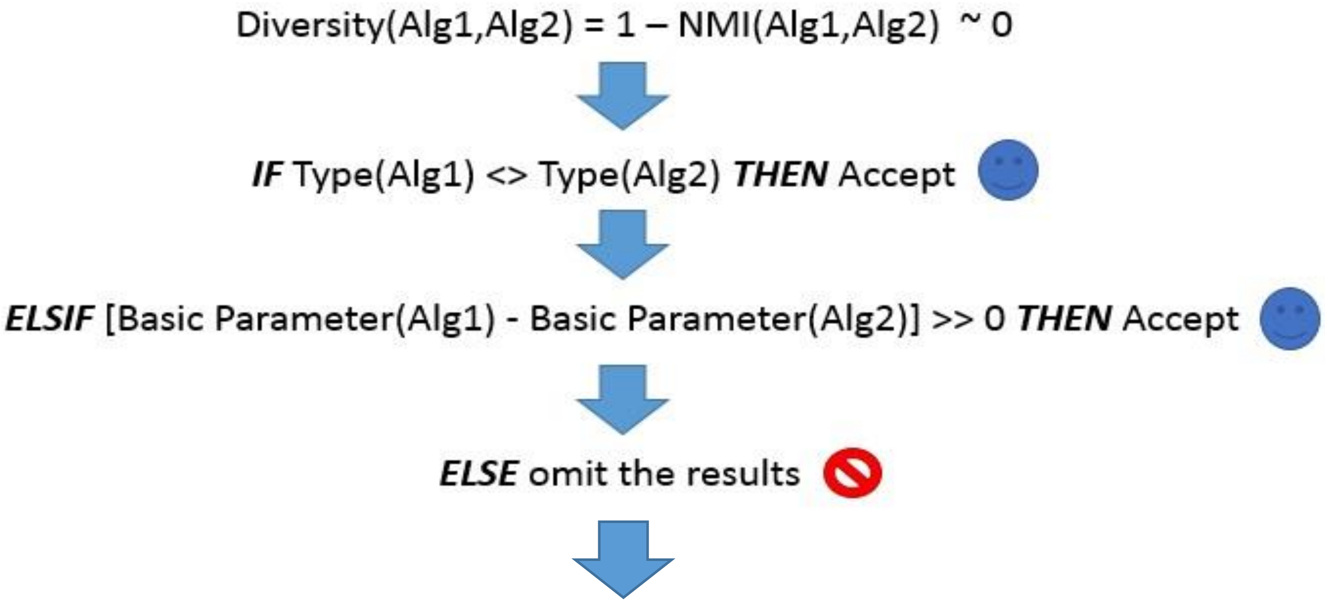
$$A3(P) = \frac{1}{n}\sum_{i=1}^{k} n_i \times AAPMM(P_i)$$

$$Diversity(P) = 1 - A3(P)$$

Diversity Condition ➡ $$Diversity(P) \geq dT$$

WOCCE – Independency

Data	Alg1	Alg2	Class
X1	1	1	1
X2	2	2	2
X3	1	1	1
X4	2	1	1
X5	2	2	2
X6	2	2	2
X7	1	1	1



Diversity(Alg1,Alg2) = 1 – NMI(Alg1,Alg2) ~ 0

IF Type(Alg1) <> Type(Alg2) *THEN* Accept ☺

ELSIF [Basic Parameter(Alg1) - Basic Parameter(Alg2)] >> 0 *THEN* Accept ☺

ELSE omit the results 🚫

Independency Condition →

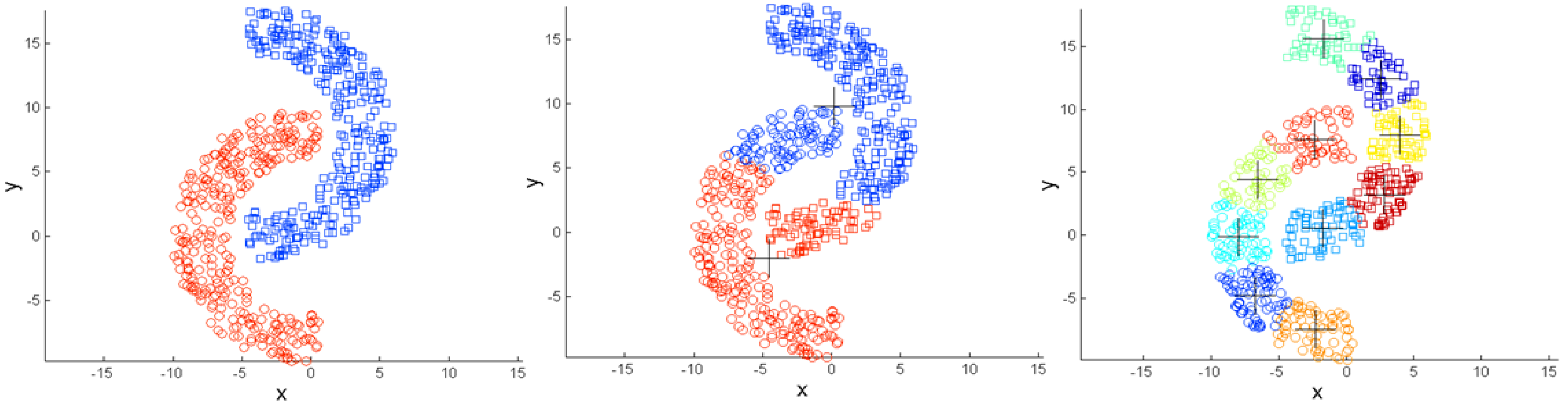
$$Independence(A\lg_i) \geq iT$$

$$Independence(P) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{BPI}(P, P_i)$$

```
Function BPI (P1, P2) Return Result
    If (Algorithm-Type (P1) == Algorithm-Type (P2) then
        Result = 1 - Likeness (Basic-Parameter (P1), Basic-Parameter (P2))
    Else
        Result = 1
    End if
End Function
```

$$Likeness = 1 - \left(\frac{1}{MaxDis}\sum_{t=0}^{n} Sim_t\right)$$

WOCCE – Decentralization



(a) Original Point (b) 2-means Clustering (c) 10-means Clustering

❑ Decentralization, a quality metric

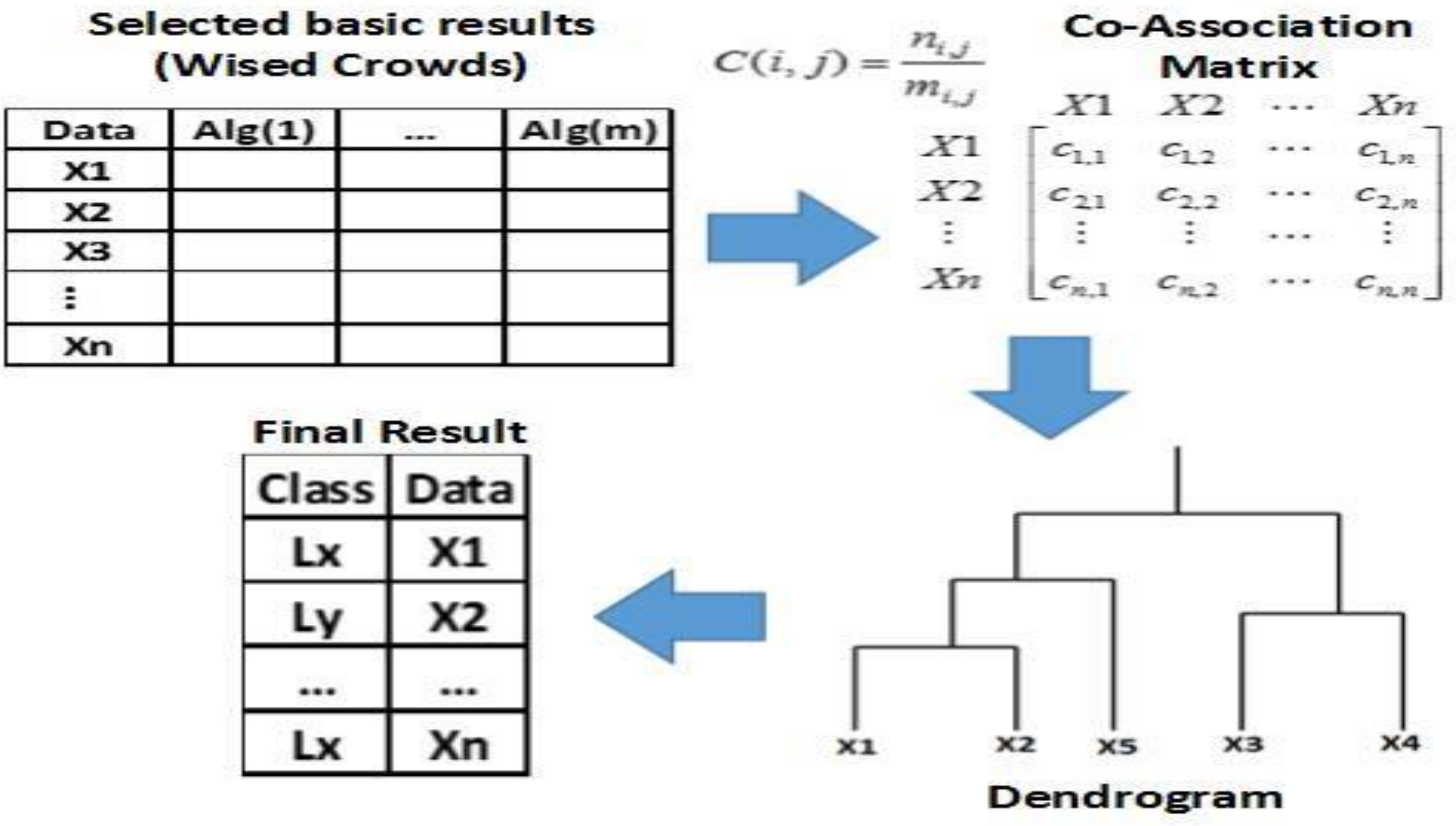
○ The number of basic algorithms participating in the crowd should be greater than one.

○ The method of entering a primary algorithm into the crowd should ensure that the final result will not be affected by its errors. In other words, the decision making of the final ensemble should not be centralized.

○ The threshold parameter **cT**, which we call the coefficient of decentralization, is a coefficient which is multiplied in the number of clusters. Every base algorithm clusters the dataset into at most **cT×k** clusters. i.e. it clusters the dataset into a number of clusters between **cT** to **cT×k**.

WOCCE – Aggregation

- **Evidence Accumulation Clustering (EAC)**



Selected basic results (Wised Crowds)

Data	Alg(1)	...	Alg(m)
X1			
X2			
X3			
⋮			
Xn			

$$C(i,j) = \frac{n_{i,j}}{m_{i,j}}$$

Co-Association Matrix

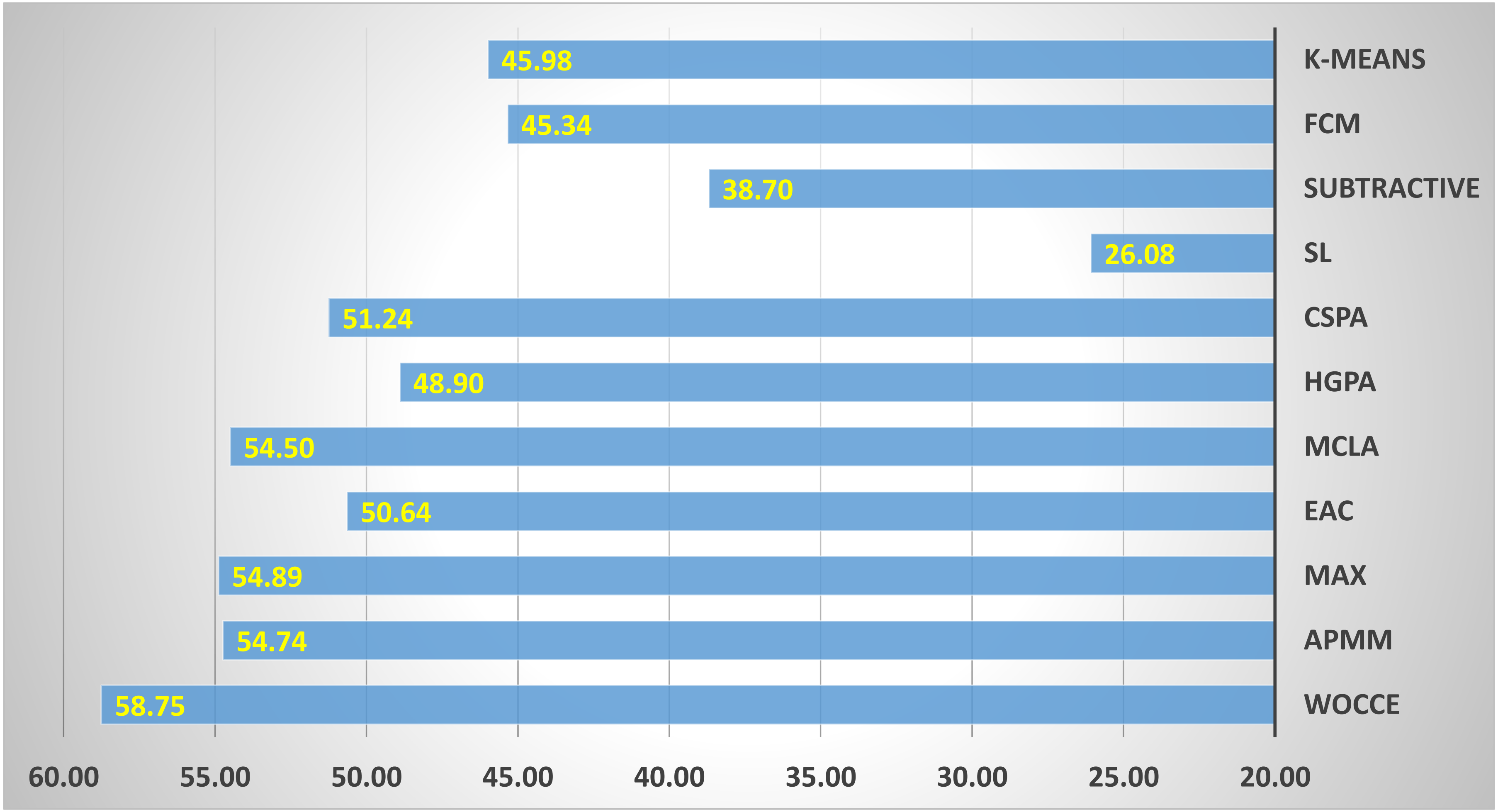
$$\begin{array}{c|cccc} & X1 & X2 & \cdots & Xn \\ \hline X1 & c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ X2 & c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ Xn & c_{n,1} & c_{n,2} & \cdots & c_{n,n} \end{array}$$

Final Result

Class	Data
Lx	X1
Ly	X2
...	...
Lx	Xn

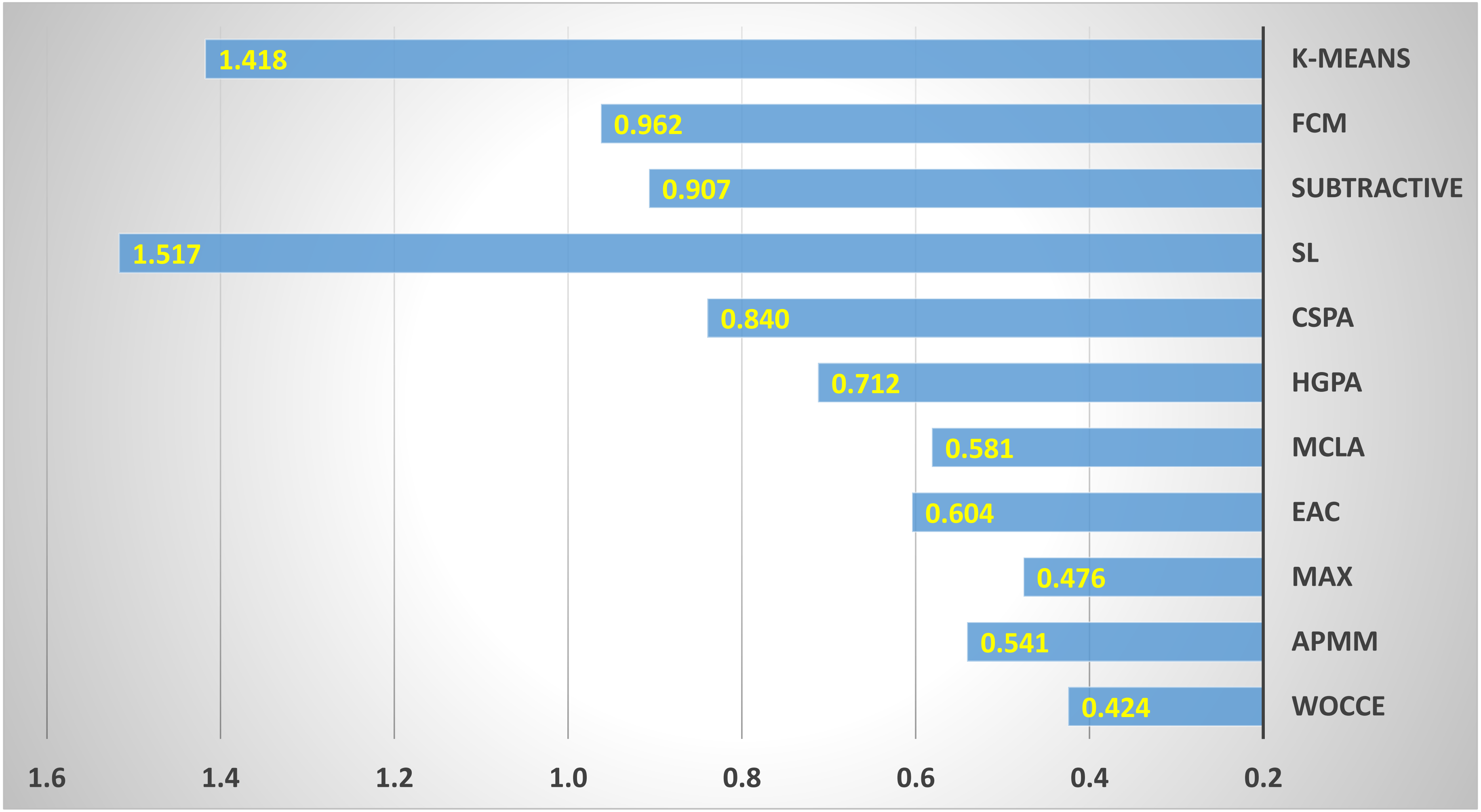
Experimental Results (Performance)

- Average of Performance: Calculated based on 10-times independence runtime on 14 standard data sets from UCI



Experimental Results (Standard error)

- Average of Error: Calculated based on 10-times independence runtime on 14 standard data sets from UCI



Feature Works

❑ Publications

- **M. Yousefnezhad**, H. Alizadeh, B. Minaei-Bidgoli, (2013), "New cluster ensemble selection method based on Diversity and Independent metrics", The 5th Conference on Information and Knowledge Technology (IKT'13), Shiraz, Iran, May 22-24/2013, Index by I.S.C (in Persian)
- H. Alizadeh, **M. Yousefnezhad**, B. Minaei-Bidgoli, (2015), "Wisdom of Crowds Cluster Ensemble", Intelligent Data Analysis, IOS Press, Vol. 19(3).
- **M. Yousefnezhad**, A. Reihanian, B. Minaei-Bidgoli, "A new selection strategy for selective cluster ensemble based on Diversity and Independency", Submitted in Apply Soft Computing (ASOC).
- **M. Yousefnezhad**, D. Zhang, A. Reihanian, "A new framework for Cluster Ensemble Selection based on Diversity and Graph based Independency", Submitted in PA-KDD 2015.
- **M. Yousefnezhad**, D. Zhang, "Wised Semi-Supervised Cluster Ensemble Selection: a new framework for selecting and combing multiple partitions based on prior knowledge", Submitted in IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI).

❑ Other ideas

- Introducing mathematician concept of Independency in CES
  - **It can be proved by probability concept in statistics**
  - **It can be proved by learning automata**
- Analyzing the effects of other basic algorithms in WOCCE framework
- Analyzing the effects of feedback mechanism in other CES techniques
- Developing WOCCE for semi-supervised learning applications

The Wisdom of Crowds Cluster Ensemble Selection

Thank You

Q&A