Fuzzy V-Measure – An Evaluation Method for Cluster Analyses of Ambiguous Data

University of Stuttgart
Germany

Jason Utt, Sylvia Springorum, Maximilian Köper, Sabine Schulte im Walde
[uttjn|riestesa|koepermn|schulte]@ims.uni-stuttgart.de

Motivation

Ambiguity is ubiquitous in language and thus methods for dealing with ambiguous data are essential for robust systems and accurate representations in NLP. Soft clusterings are for instance a very natural strategy for representing **ambiguous data**. However, the **evaluation** methods are still missing a suitable **measure for comparing the soft cluster analyses**. This work aims to fill this gap.

Fuzzy V-Measure

Because of fuzzy data (data points can belong to multiple clusters, which means that clusters are not disjoint), **joint probability with** simple intersection $|c \cap g|$ with normalising constant N doesn't work. Therefore we use a **mass function** μ:

$$\hat{p}(c,g) = \frac{\mu(c \cap g)}{M}$$

μ: Total mass of the objects in the data shared by c and g
M: Total mass of the clustering

→ Each data point is assigned with a total mass of *1* and then evenly distributed among its classes and then normalised with the total mass of the clustering

V- Measure (Rosenberg and Hirschberg, 2007)

Measure for comparison of two completely independent clusterings with no restrictions in their similarity, the number of data points, or the number of clusters.

$$v_\beta(C) = \frac{(1+\beta) \cdot hom(C) \cdot com(C)}{\beta \cdot hom(C) + com(C)}$$

→ A **weighted harmonic mean of homogeneity and completeness** values.

Homogeneity
Measure of how homogeneous the clusters in the clustering are

$$hom(C) = \begin{cases} 1 & \text{if } H(C,G) = 0; \\ 1 - \frac{H(C|G)}{H(C,G)} & \text{else} \end{cases}$$

Completeness
Measure of how intact the gold standard classes remain with respect to the clustering

$$com(C) = \begin{cases} 1 & \text{if } H(G,C) = 0 \\ 1 - \frac{H(G|C)}{H(G,C)} & \text{else} \end{cases}$$

H(C|G): Conditional entropy of *C* given *G*
H(G|C): Conditional entropy of *G* given *C*

H(C,G) and *H(G,C)*: Joint entropies for normalisation

Entropies are calculated with the **joint probability** of a cluster and a gold standard class.

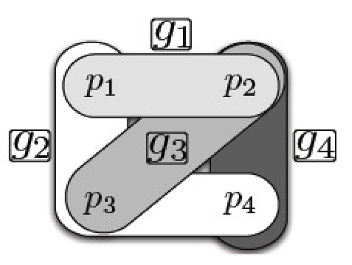
$$\hat{p}(c,g) = \frac{|c \cap g|}{N}$$

The number of **points shared by *c* and *g*** divided by the total number of data points *N*.

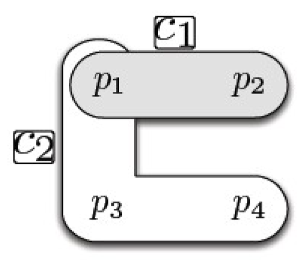
Example

Data points: p_1, p_2, p_3, p_4
Gold standard classes: g_1, g_2, g_3, g_4
Clusters: c_1, c_2

Distribution of ambiguous data points:



Clustering of ambiguous data:



E.g. V-Measure would assign the pair p_2 and g_4 and p_4 and g_4 the same joint probability, but p_2 belongs to three classes and p_4 to two
→ Too much weight on highly ambiguous objects

(a) Distribution of data points in gold standard with mass of objects (*1* divided number of *g*)

(b) Contingency table containing mutual evidence between classes and cluster based on the new, above introduced object distribution using the adjusted joint probability with mass function μ.
(c.f. section Fuzzy V-Measure)

	g_1	g_2	g_3	g_4
p_1	.5	.5	0	0
p_2	.33	0	.33	.33
p_3	0	.5	.5	0
p_4	0	.5	0	.5

(a)

	g_1	g_2	g_3	g_4	Σ
c_1	.83	.5	.33	.33	= 2
c_2	.5	1.5	.5	.5	= 3

(b)

Advantages of Fuzzy V:

c_1 and g_1 share points p_1 and p_2
c_2 and g_2 share points p_1, p_3 and p_4

The highly ambiguous p_2 reduces the evidence for: c_1 given g_1 $p(c_1|g_1) = 0.83/2$

Even though c_1 and c_2 share all objects, the evidence is smaller than for: c_2 given g_2 $p(c_2|g_2) = 1.5/3 = 1/2$

→ Incorporates ambiguity of data points

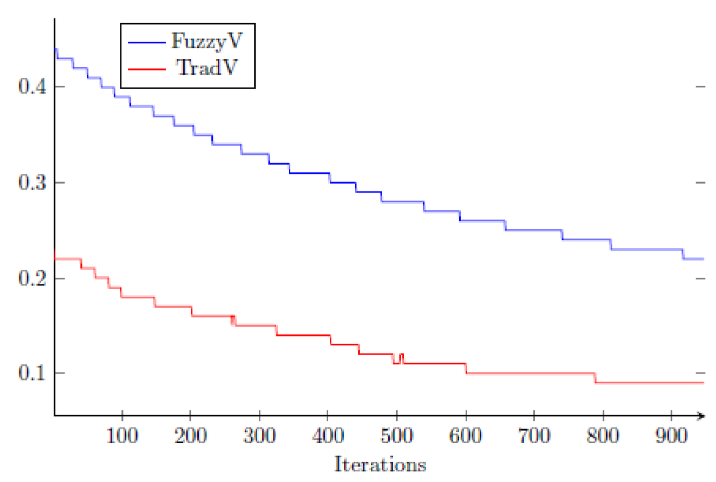
Applying V-Measures

Using different data settings:

Experiment 1:
Data sets varying in the amount of different ambiguity rates of the objects assuming a perfect clustering
→ None of the measures reach the expected value of *1*

Experiment 2:
Comparing arbitrary clusterings with random objects with constant ambiguity rate across different data sizes.
→ Fuzzy V is less sensitive to ambiguity than V

Experiment 3:
Comparing variation in the ambiguity rate while maintaining the data points
→ Both values decrease with each cluster closer to the fuzzy gold stand.



x-axis: stepwise increase of ambiguity rate

Beyond Entropy

Why does the curve decrease the more ambiguous a data set is?
Because of the entropy: Increased spread of mass (due to the ambiguity) leads to an increase in the overall uncertainty in the correspondence between clusters and classes

Why do the perfect clusterings not reach the maximum score 1?
Example table for perfect hard clusterings (a) and (b) vs. table for perfect soft clustering (c):

$$\begin{matrix} & g_1 & g_2 & g_3 \\ c_1 & 2 & 0 & 0 \\ c_2 & 0 & 2 & 0 \\ c_2 & 0 & 0 & 2 \end{matrix} \qquad \begin{matrix} & g_1 & g_2 & g_3 \\ c_1 & 0 & 2 & 0 \\ c_2 & 0 & 0 & 2 \\ c_2 & 2 & 0 & 0 \end{matrix} \qquad \begin{matrix} & g_1 & g_2 & g_3 \\ c_1 & \boxed{1} & 2 & 0 \\ c_2 & 2 & \boxed{1} & 0 \\ c_2 & 0 & 0 & 2 \end{matrix}$$

(a) (b) (c)

E.g. in (c) g_1 and g_2 share one ambiguous element which lead to similarity between them and to double entries between several cluster/gold-class pairings.
→ Score less than *1*

We propose to include **Dissimilarity,** which enables us to

1. Force a one-to-one mapping between c_x and g_x with high similarity and low dissimilarity
2. Penalise other mappings, by uniformly distributing the remaining error mass (e_x is the dissimilarity between the best mapping c_x and g_x)

Similarity: Shared elements' mass
Dissimilarity: Missing and remaining elements between all cluster/class combinations

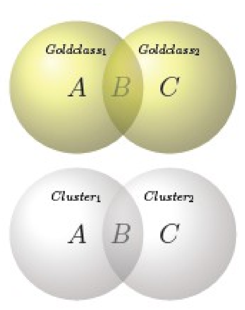
Example 1: 3 elements *A,B,C; B* is ambiguous; perfect Clustering

$$sim = \begin{matrix} & g_1 & g_2 \\ c_1 & 2 & 1 \\ c_2 & 1 & 2 \end{matrix}$$

(a) 'Hard' contingency table

$$sim = \begin{matrix} & g_1 & g_2 \\ c_1 & 1.5 & 0.5 \\ c_2 & 0.5 & 1.5 \end{matrix}$$

(b) 'Soft' contingency table



$$diss = \begin{matrix} & g_1 & g_2 \\ c_1 & 0 & 1 \\ c_2 & 1 & 0 \end{matrix}$$

(a) Hard dissimilarity matrix

$$diss = \begin{matrix} & g_1 & g_2 \\ c_1 & 0 & 0.5 \\ c_2 & 0.5 & 0 \end{matrix}$$

(b) Soft dissimilarity matrix

Decision for final cluster to class mapping with the highest score of **difference between Similarity and Dissimilarity**.

$$score = \begin{matrix} & g_1 & g_2 \\ c_1 & \boxed{2} & 0 \\ c_2 & 0 & \boxed{2} \end{matrix}$$

(a) hard score

$$score = \begin{matrix} & g_1 & g_2 \\ c_1 & \boxed{1.5} & 0 \\ c_2 & 0. & \boxed{1.5} \end{matrix}$$

(b) soft score

Resulting mapping:
$c_1 \rightarrow g_1$ and $c_2 \rightarrow g_2$

Error mass = Dissimilarity value for the best mapping: *0*

Example 2: Clustering and gold standard are different
if error mass > *0*, it will be distributed equally among the non-zero entries in each row

Goldclass: $g_1 : A, B$,$g_2 : B, C$
Cluster: $c_1 : A, B$,$c_2 : C$

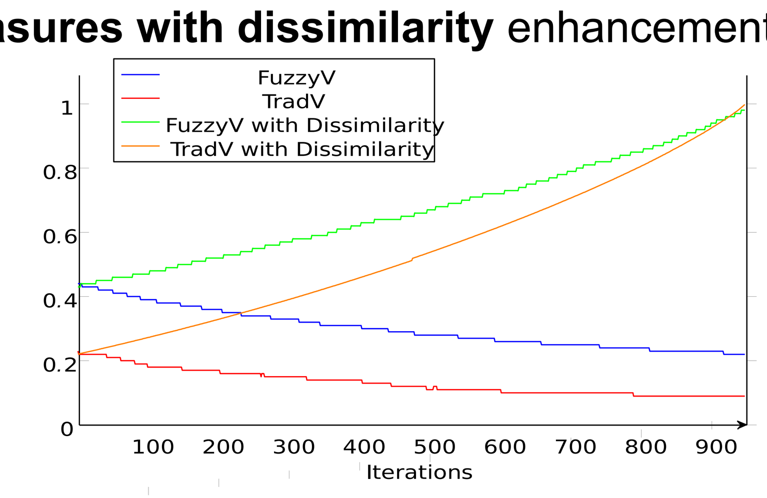
$$sim = \begin{matrix} & g_1 & g_2 \\ c_1 & 1.5 & 0.5 \\ c_2 & 0 & 1 \end{matrix} \quad diss = \begin{matrix} & g_1 & g_2 \\ c_1 & 0 & 2 \\ c_2 & 2 & 0.5 \end{matrix}$$

$$score = \begin{matrix} & g_1 & g_2 \\ c_1 & \boxed{1.5} & -1.5 \\ c_2 & -2 & \boxed{0.5} \end{matrix}$$ highest value determines assignment

$$cont = \begin{matrix} & g_1 & g_2 \\ c_1 & 1.5 & 0 \\ c_2 & 0.5 & 0.5 \end{matrix}$$ Errormass for $c_1 = 0$
Errormass for $c_2 = 0.5$

Performance of the **V-Measures with dissimilarity** enhancement:

- for perfect clustering
- with stepwise increase of ambiguity rate (*x*-axis)



→ Both measures converge toward the desired score of *1*

Conclusion: A purely entropy based measure cannot capture the complexity of highly ambiguous data sets.
A further disambiguation, e.g. **Dissimilarity difference,** on the cluster/class assignment is required.

Reference

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Learning (EMNLP-CoNLL)*, pages 410-420