V-Measure: A conditional entropy-based external cluster evaluation measure.

Andrew Rosenberg and Julia Hirschberg
Department of Computer Sciente
Columbia University
New York, NY 10027
{amaxwell,julia}@cs.columbia.edu

Abstract

In this paper we present V-measure, an external entropy-based cluster evaluation metric. V-measure provides an elegant solution to many problems that affect previously defined cluster evaluation measures including 1) dependence on clustering algorithm or data set, 2) accurate evaluation and combination of two aspects of good clustering - homogeneity and completeness, and 3) symmetry in the measurement of these for easy interpretation. We draw comparisons between V-measure and a number of popular cluster evaluation measures, as well as empirically showing that V-measure satisfies a number of desirable properties of clustering solutions based on a simulated clustering results. We present the use of V-measure to evaluate two example clustering tasks: document clustering and pitch accent type clustering.

1 Introduction

Clustering techniques are particularly appealing for many natural language processing tasks.

However, evaluating the "goodness" of a clustering solution is a critical and difficult empirical problem (**?**) and often lacks rigor (**?**).

There are two criteria of a successful clustering solution. First, the homogeneity criteria: each cluster should contain only data points that are members of a single class. Second, the completeness criteria: all of the data points that are members of a given class should be elements of the same cluster. We believe that any external[1] metric for evaluating a clustering solution should determine to what degree both of these criteria are satisfied.

The criteria of homogenity and completeness are roughly in opposition; increasing the homogeneity of a clustering solution often results in a decrease of completeness. Consider, wo degenerate clustering solutions. One, assigning every datapoint into a single cluster guarantees perfect completeness – all of the data points that are members of the same class are trivially elements of the same cluster. However, this cluster is as far from homogeneous as possible – the maximum diversity of classes are represented in this single cluster. Two, assigning each data point to a distinct cluster guarantees perfect homogeneity – each cluster trivially only contains members of a single class. However, by virtue of each cluster containing only a single member of a class the clustering solution is as far from complete as possible.

In this paper, we present a new external cluster evaluation metric, V-measure. V-measure is an entropy-based metric which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores. This is identical to the way precision and recall are commonly combined into F-measure(**?**) for evaluation of information retrieval results. Just as F-measure scores can be weighted to favor the contributions of precision or recall, V-measure can be weighted to favor the contributions

[1] Obviously, if the class labels of the datapoints are not known *a prori*, these criteria cannot be applied

of homogeneity and completeness.

In Section 2 we present the calculation of V-measure. We discuss some popular external cluster evaluation metrics and draw comparisons between these and V-measure in Section 3. We present some empirical desirable properties and describe the degree to which are satisfied by V-measure and other measures in section **??**. In Section 5 we will show two examples of the application of V-measure on two clustering tasks: document clustering, and pitch accent clustering.

2 Calculating V-Measure

In order to use an external clustering metric, class labels for each data point, or a reference partition must be known a priori. The clustering task is to assign these data points to any number of clusters such that each cluster contains all and only those data points that are members of the same class. Given the ground truth class labels, it is trivial to determine if this perfect clustering has been achieved. However, evaluating how far from perfect an incorrect clustering solution is a more difficult task. We propose to measure this distance from perfection as the weighted harmonic mean of two measures, evalutaing the degree to which the homogeneity and completeness criteria described in Section 1 have been satisfied.

For the purposes of the following discussion, assume N data points, a set of classes, $C = \{c_i | i = 1, \ldots, n\}$ and a set of clusters, $K = \{k_i | 1, \ldots, m\}$. Let A be the contingency table produced by the clustering algorithm, such that $A = a_{ij}$ such that a_{ij} is the number of data points that are members of class c_i and elements of cluster k_j.

Homogeneity:

In order to satisfy the homogeneity criteria, a clustering must assign **only** those datapoints that are members of a single class to a single cluster. That is the class distribution within each cluster should be totally skewed to a single class, that is, zero entropy. We determine how close a given clustering is to this ideal by examining the conditional entropy of the class distribution given the proposed clustering. In the perfectly homogeneous case, this value, $H(C|K) = 0$. However, in an imperfect situation the size of this value, in bits, is dependent on the size of the dataset. Therefore, instead of taking the raw conditional entropy, we normalize this value by the maximum reduction in entropy the clustering information could provide, specifically, $H(C, K)$. Technically, this is a weak upper bound, $H(C|K) \leq H(C) \leq H(C, K)$, however, normalization by $H(C)$ yields a measure that behaves in unintuitive, and undesirable ways.

Note that $H(C|K)$ is maximal when the clustering provided no new information – the class distribution within each cluster is even, and 0, when each cluster contains only members of a single class, a perfectly homogenous clustering. In the degenerate case where $H(C, K) = 0$ we defined both homogeneity and completeness to be 1. For a perfectly homogenous solution the normalization $H(C|K)/H(C, K)$ equals 0. Thus, to adhere to the convention of 1 being 'good' and 0 'bad', we define homogeneity as:

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

where

$$H(C|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{A_{ck}}{N} \log \frac{A_{ck}}{\sum_{c=1}^{|C|} A_{ck}}$$

$$H(C, K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{A_{ck}}{N} \log \frac{A_{ck}}{N}$$

Completeness:

Completeness is calculated symmetrically to homogeneity. In order to satisfy the completeness criteria, a clustering must assign **all** of those datapoints that are members of a single class to a single cluster. To evaluate this we examine the distribution of cluster assignments within each class. In a perfectly complete clustering, each of these distributions will be completely skewed to a single cluster. Similar to the above, we can evaluate this by calculating the conditional entropy of the proposed cluster distribution given the class of the component datapoint, $H(K|C)$. In the perfectly complete case, $H(K|C) = 0$. However, in the worst case scenario, each class is equally represented by every cluster

where $H(K|C)$ is maximal. Therefore, symmetrically to the calculation above, we define completeness as:

$$c = \begin{cases} 1 & \text{if } H(K,C) = 0 \\ 1 - \frac{H(K|C)}{H(K,C)} & \text{else} \end{cases} \quad (2)$$

where

$$\begin{aligned} H(K|C) &= -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{A_{ck}}{N}\log\frac{A_{ck}}{\sum_{k=1}^{|K|}A_{ck}} \\ H(K,C) &= H(C,K) \end{aligned}$$

Finally, we calculate V-measure by by computing the weighted harmonic mean of homogeneity and completeness. Similarly with F-measure, if β is greater than 1 completeness is weighted more strongly in the calculation, if β is less than 1, homogeneity is weighted more strongly.

$$V_\beta = \frac{(1+\beta) * h * c}{(\beta * h) + c} \quad (3)$$

Notice that the computations of homogeneity, completeness and V-measure are completely independent of the number of classes, the number of clusters, the size of the data set and the clustering algorithm used. This allows these metrics to be applied to and compared across any clustering solution, regardless of the number of data points (n-invariance), the number of classes or the number of clusters. Moreover, by calculating homogeneity and completeness separately, a more precise evaluation of the performance of the clustering is offered.

3 Existing External Clustering Evaluation Techniques

There exist are a number of proposed external clustering measures. However, we find these to be lacking in a variety of ways.

Two commonly used metrics are Purity and Entropy (**?**), defined as,

$$Purity = \sum_{r=1}^{k}\frac{1}{n}\max_i(n_r^i) \quad (4)$$

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}(-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}) \quad (5)$$

where q is the number of classes, k the number of clusters, n_r is the size of cluster r, and n_r^i is the number of data points in class i clustered in cluster r.

Both Purity and Entropy plausable ways to evaluate the homogeneity of a clustering solution, however the completeness criterion is not measured at all by these metrics. Therefore the Purity and Entropy are likely to improve (increased Purity, decreased Entropy) monotonically with the number of clusters used, up to a degenerate maximum where there are as many clusters as data points. This is certainly not an ideal clustering, depite yielding high Purity and low Entropy scores.

Another common external clustering evaluation metric is generally refered to as "clustering accuracy". The calculation of "accuracy" is inspired by the information retrieval metric of F-Measure (**?**). The formula for this clustering F-measure as described in (**?**) can be found in Figure 3.

Let N be the number of data points, C the set of classes, K the set of clusters and n_{ij} be the number of members of class $c_i \in C$ that are elements of cluster $k_j \in K$.

$$F(C,K) = \sum_{c_i \in C}\frac{|c_i|}{N}\max_{k_j \in K}\{F(c_i,k_j)\} \quad (6)$$

$$F(c_i,k_j) = \frac{2 * R(c_i,k_j) * P(c_i,k_j)}{R(c_i,k_j) + P(c_i,k_j)}$$

$$R(c_i,k_j) = \frac{n_{ij}}{|c_i|}$$

$$P(c_i,k_j) = \frac{n_{ij}}{|k_j|}$$

Figure 1: Calculation of clustering F-measure

This metric has a significant advantage over Purity and Entropy. Specifically, this is the capability to measure both the homogeneity and completeness of a clustering solution. Recall, equation (x), is calculated as the portion of items from class i that are present in cluster j. This is essentially a measure of how complete cluster j is with respect to class i. Similarly, Precision, equation (y), is calculated as the portion of cluster j that is a member of class i. This measures how homogenous cluster j is with

respect to class i.

F-measure is a member of a class of external cluster evaluation functions that rely on a post-processing step in which each cluster is assigned to a class. These metrics include misclassification index (MI) (**?**), H (**?**), L (**?**), D (**?**). There are two major problems with these metrics. First of all, they calculate the goodness not only of the given clustering solution, but also the cluster-class matching. Therefore, in order for the goodness of two clustering solutions to be compared using one these metrics, an identical post-processing algorithm must be used. This problem can be trivially addressed by fixing the class-cluster matching function and including it in the definition of the measure as in H. A second more critical problem is the "problem of matching" (**?**). In calculating the similarity between a hypothesized clustering and a 'true' clustering, these measures only consider the contributions from those clusters that are matched to a true set. The unmatched portion of each cluster is not evaluated in the distance function.

For example, consider figure **??**. In both clustering solutions, 4/7 of each cluster is 'matched' to a true cluster. However, the make of the unshaded regions of each cluster are not included in any of these measures. Arguments regarding whether solution A or B is "better" can be deferred for the moment, regardless, it is clear that these two clusterings are considerably different. A clustering evaluation function should measure this difference. Each of MI, H, L and F-measure consider these solutions to be equivalent with respect to the target clustering.

A second class of cluster evaluation techniques are based on combinatorial approach which examines the number of pairs of data points that are clustered similarly in the target and hypothesized clustering. That is, each pair of points can either 1) clustered together in both clusterings, 2) clustered separately in both clusterings, 3) clustered together in the hypothesized but not the target clustering or 4) clustered together in the target but not in the hypothesized clustering. Based on these 4 values, a number of measures have been proposed including Rand Index (**?**), Adjusted Rand Index (**?**), Γ statistic (**?**), Jacard (**?**), Fowlkes-Mallows (**?**) and Merkin (**?**). We reproduce the calculation of Rand Index in Figure 3 as an example of this class of evaluation measures.

Let N_{11} be the number of pairs of data points that are clustered together in both the target (C) and hypothesized (K) clusterings. Let N_{00} be the number of pairs clustered separately in both C and K. Let n be the size of the data set.

$$R(C,K) = \frac{N_{11} + N_{00}}{n(n-1)/2} \quad (7)$$

Rand Index can be interpreted as the probability that a pair of points is clustered similarly (together or separately) in C and K.

Figure 2: Calculation of Rand Index

Meila (**?**) reiterates a number of potential problems afflicting this class of measures posed by both (**?**) and (**?**). The most trivial of these problems is that these metrics tend not to vary over the interval of $[0, 1]$. Transformations like those applied by the adjusted rand index and a minor adjustment to the Merkin metric (see Section **??**) are able to sidestep this problem. However, there is also a distributional problem. The baseline for Fowlkes-Mallows varies significantly between 0.6 and 0 when the ratio of data points to clusters is greater than 3 (obviously including nearly all real-world clustering problems). Similarly, the Adjusted Rand Index, as demonstrated using Monte Carlo simulations in (**?**) varies from 0.5 to 0.95. This variance in the metric's baseline prompts Meila to ask if the assumption of linearity following normalization can be maintained. That is, if the behavior of the metric is so unstable before normalization can users responsibly expect stable behavior following normalization?

A final class of cluster evaluation measures are based on information theory. These measures analyze the distribution of class and cluster memberships in order to determine how successful a given clustering solution is or how different two partitions of a data set are. We have already examined one member of this class of measures, $Entropy$. From a coding theory perspective, $Entropy$ is the weighted average of the code lengths of each cluster. V-measure (see Section 2) is also a member of this class.

One significant advantage that information theo-

retic evaluation measures have is that they provide an elegant solution to the "problem of matching". By examining the relative sizes of the classes and clusters being evaluated, these measures all evaluate the entire membership of each cluster – not only a 'matched' portion.

The Q_0 measure described in (**?**) uses conditional entropy, $H(C|K)$ to calculate the goodness of a clustering solution. That is, given the hypothesized partition, what is number of bits necessary to represent the true clustering. If $C = K$, $H(C|K) = 0$. However, this term – like the $Purity$ and $Entropy$ measures – only evaluates the homogeneity of a solution. To account for the completeness of the hypothesized clustering, Dom includes a model cost term calculated using a coding theory argument. The clustering quality measure presented is then the cost of representing the data ($H(C|K)$) and the cost of representing the model. The motivation for this is an appeal to parsimony; given identical conditional entropies, $H(C|K)$, the clustering solution with the fewest clusters should be preferred.

$$Q_0(C,K) = H(C|K) + \frac{1}{n}\sum_{k=1}^{|K|} \log \binom{h(k)+|C|-1}{|C|-1} \quad (8)$$

We believe V-measure provides two significant advantages that allow for it to serve as a more useful diagnostic tool than Q_0. First of all, Q_0 does not explicitly calculate the degree of completeness of the clustering algorithm. The cost term captures some of this information because a partition with fewer clusters is likely to be more complete than a clustering solution with more clusters. However, Q_0 does not explicitly address the interaction between the conditional entropy and the cost of representing the model. While this is an application of the *minimum description length*(MDL) principle (**?**; **?**), it does not provide a intuitive manner for assessing the two competing criteria of homogeneity and completeness. That is, at what point does an increase of conditional entropy (homogeneity) justify a reduction in the number of clusters (completeness).

Another information based clustering measure is variation of information (VI) (**?**). VI is presented as a distance metric for comparing partitions (or clusterings) of the same data. VI, therefore, does not distinguish between hypothesized and target clusterings.

$$VI(C,K) = H(C|K) + H(K|C) \quad (9)$$

VI has a number of very useful properties. First of all, VI satisfies the metric axioms. Specifically, 1) it is always non-negative and only equals zero of $C = K$, 2) it is symmetric, 3) the triangle inequality holds. This quality allows users to intuitively understand how VI values combine and relate to one another. Secondly, it is "convexly additive". That is to say, if a cluster is split, the distance from the new cluster to the original is the distance induced by the split times the size of the cluster. This property guarantees that all changes to the metric are "local": the impact of splitting or merging clusters is affected only by those clusters involved, and its size is relative to the size of these clusters.

Another property of VI is that it is n-invariant: the number of data points in the cluster do not affect the value of the measure. VI depends on the relative sizes of the partitions of C and K, not on the number of points in these partitions. However, VI is bound by the maximum number of clusters in C or K, k^*. Without manual modification however, $k^* = n$, where each cluster contains only a single data point. Thus, while technically n-invariant, the possible values of VI are very influenced by the number of data points being clustered. This makes it difficult to compare VI values across data sets and clustering algorithms without fixing k^* as VI will vary over different ranges. However, it is a trivial modification to modify VI such that it varies over [0,1]. Normalizing, VI by $\log n$ or $1/2 \log k^*$ guarantee this range. Meila (**?**) raises two potential problems with these. First of all, the former normalization shouldn't be applied if data sets of different sizes are to be compared – it negates the n-invariance of the metric. Secondly, if two authors apply the latter normalization and do not use the same value for k^*, their results will not be comparable. Moreover, where VI is always measured in bits, these normalizations are measured in arbitrary units.

While VI demonstrates a number of useful numeric properties, these last encumbers its application in comparing results across disparate cluster-

ings of disparate data sets. Homogeneity (h) and completeness (c) as described in section 2 both range over [0,1] and are completely n-invariant and k^*-invariant. Regarding the measurement unit of h and c, they are each measured as a ratio of bit lengths, while this is technically an 'arbitrary' unit, it has greater intuitive appeal than a more opportunistic normalization. While VI has a number of very useful distance properties when analyzing a single data set across a number of settings, we believe the impact of n or k^* to limit its usefulness as a general purpose clustering evaluation metric.

V-measure has another advantage as a clustering evaluation measure over VI and Q_0. By evaluating homogeneity and completeness in a symmetrical, complementary manner, the calculation of V-measure makes their relationship clearly observable. Separate analyses of homogeneity and completeness are not possible with any other cluster evaluation measure. Moreover, by using the harmonic mean to combine homogeneity and completeness, V-measure can be made sensitive to priorities of one criteria over another depending on the clustering task and goals. Similar sensitivity is not possible with Q_0 or VI. While this sacrifices any possibility of satisfying the metric axioms in its general form, we don't believe that a cluster evaluation measure should necessarily be symmetric. Knowledge of which partitioning is the target and which is hypothesized allows insight into not only "how similar" the two are, but also "in what way".

4 Desirable Properties

Dom (**?**) describes a parametric technique for generating example clustering solutions. He then proceeds to define five "desirable properties" that clustering accuracy measures should display, based on the parameters used to generate the clustering solution. We evaluate V-measure against these and two additional desirable properties.

The parameters used in generating a clustering solution are as follows.

- $|C|$ The number of classes
- $|K|$ The number of clusters
- $|K_{noise}|$ Number of "noisy" clusters; $|K_{noise}| < |K|$
- $C_{noise}|$ Number of "noisy" classes; $|C_{noise}| < |C|$
- ϵ Error probability; $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3$.
- ϵ_1 The error mass within "useful" class-cluster pairs
- ϵ_2 The error mass within noisy clusters
- ϵ_3 The error mass within noisy classes

The construction of a clustering solution begins with a matching of "useful" clusters to "useful" classes[2]. There are $|K_u| = |K| - |K_noise|$ "useful" clusters and $|C_u| = |C| - |C_{noise}|$ "useful" classes. Probability mass of $1-\epsilon$ is evenly distributed across each match. Error mass of ϵ_1 is evenly distributed across each pair of non-matching useful class/cluster pairs. Error mass of ϵ_2 is distributed across every "noise"-cluster/ "useful"-class pair. Error mass of ϵ_3 is distributed across every cluster/"noise"-class pair. An example solution, along with its generating parameters is given in Figure 3.

	C_1	C_2	C_3	C_{noise1}
K_1	33	33	6	9
K_2	6	6	33	9
K_{noise1}	12	12	12	9

Figure 3: Sample parametric clustering solution with $|K| = 3, |K_{noise}| = 1, |C| = 3, |C_{noise}| = 1, \epsilon_1 = .1, \epsilon_2 = .2, \epsilon_3 = .15$

The desirable properties proposed by Dom are P1-P5 in Table 1. Dom did not include the parameter and error term for "noise" classes, therefore P6, P7 were not evaluated in (**?**).

We systematically varied each parameter keeping $|C| = 5$ fixed.

- $|K_u|$: 10 values: 2, 3,. . . , 11
- $|K_{noise}|$: 7 values: 0, 1,. . . , 6
- $|C_{noise}|$: 7 values: 0, 1,. . . , 6
- ϵ_1: 4 values: 0, 0.066, 0.133, 0.2
- ϵ_2: 4 values: 0, 0.1, 0.2, 0.3
- ϵ_3: 4 values: 0, 0.1, 0.2, 0.3

[2]The operation of this matching is omitted in the interest of space. Interested readers are encouraged to refer to (**?**).

We evaluated the behavior of V-Measure, Rand, Merkin, Fowlkes-Mallows, Gamma, Jacard, VI, Q_0, Entropy, F-Measure against the desirable properties P1-P7. Based on the described systematic modification of each parameter, only V-measure, VI and Q_0 empirically satisfy all of P1-P7 in all experimental conditions.

5 Applications

5.1 Document Clustering

Clustering techniques have been used considerably in clustering documents into topic clusters. We reproduce this type of experiment here to demonstrate the use of V-measure. Using a subset of the TDT-4 corpus (1884 English news wire and broadcast news documents that were manually labeled with one of 12 topics), we ran clustering experiments using with k-means (**?**) and average-linkage hierarchical clustering (**?**). The topics and relative distributions are as follows: Acts of Violence/War (22.3%), Elections (14.4%), Diplomatic Meetings (12.9%), Accidents (8.75%), Natural Disasters (7.4%), Human Interest (6.7%), Scandals (6.5%), Legal Cases (6.4%), Miscellaneous (5.3%), Sports (4.7), New Laws (3.2%), Science and Discovery (1.4%).

We used stemmed, tf*idf-weighted term vectors extracted for each document as the clustering space for these experiments. However, this yielded a very high dimension space. In order to reduce this dimensionality, we performed a crude feature selection procedure. We included in the feature vector only those terms that represented the highest tf*idf value for at least one data point. This resulted in a feature vector containing 484 tf*idf values for each document. Results average-linkage hierarchical cluster can be seen in Figure 4. Results from both k-means and average linkage can be observed in Figure 5.

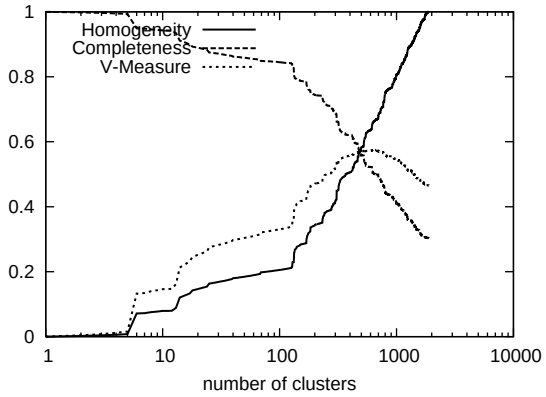

Figure 4: Results of document clustering measured by V-Measure, homogeneity and completeness

P1 For $|K_u| < |C|$ and $\Delta|K_u| \leq (|C| - |K_u|)$, $\frac{\Delta M}{\Delta|K_u|} > 0$

P2 For $|K_u| \geq |C|$, $\frac{\Delta M}{\Delta|K_u|} < 0$

P3 $\frac{\Delta M}{\Delta|K_{noise}|} < 0$

P4 $\frac{\delta M}{\delta\epsilon_1} \leq 0$, with equality only if $|K_u| = 1$

P5 $\frac{\delta M}{\delta\epsilon_2} \leq 0$, with equality only if $|K_{noise}| = 0$

P6 $\frac{\Delta M}{\Delta|C_{noise}|} < 0$

P7 $\frac{\delta M}{\delta\epsilon_3} \leq 0$, with equality only if $|C_{noise}| = 0$

Table 1: Desirable Properties of a cluster evaluation measure M

5.2 Pitch Accent Clustering

Pitch accent is how speakers of Standard American English indicate that a word in an utterance is prominent. Moreover, words can be accented in different ways to indicate different types of emphasis (**?**) and discourse structure (**?**). These different ways have been categorized into discrete "pitch accent types" by the ToBI labeling scheme (**?**). In this clustering experiment, we extract a number of acoustic features from accented words within the read portion of the Boston Directions Corpus (BDC) (**?**) and examine how well clustering in these acoustic dimensions correlates to manually annotated pitch accent types. The read portion of the BDC corpus contains read transcripts of increasingly complicated direction giving tasks. The speech is produced by four non-professional speakers (three male and one female). The transcripts that were read by each speaker, were initially produced spontaneously at an earlier session by the same speaker. We collapse all downsteped instances of pitch accents with corresponding non-downsteped instances for these experiments. This left a very skewed distribution with a majority of H pitch accents. We therefore included a randomly selected 10% sample of H* accents. This left a more even distribution (see Table 2) of pitch

accent types for clustering.

H*	L*	L+H*	L*+H	H+!H*
35.4%	32.1%	26.5%	2.8%	2.1%

Table 2: Distribution of Pitch Accent Types

We extract ten acoustic features from each accented word to serve as the clustering space for this experiment. Using Praat's (**?**) Get Pitch (ac)... function, we calculated the mean F0 and ΔF0, as well as z-score speaker normalized versions of the same. We included in the feature vector the relative location of the maximum pitch value in the word as well as the distance between this maximum and the point of maximum intensity. Finally we calculated the raw and speaker normalized slope from the start of the word to the maximum pitch, and from the maximum pitch to the end of the word.

Using this feature vector, we perform k-means clustering and average-linkage hierarchical clustering and evaluate how successfully these dimensions represent differences between pitch accent types. The results can be seen in Figure 5.

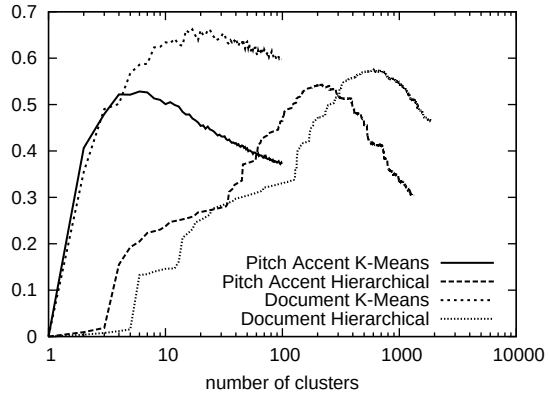

Figure 5: Results of all clustering experiments evaluated using V-Measure

5.3 Discussion

In figure 4, the relationship between homogeneity and completeness is clearly observable; as the number of clusters increase we see homogeneity increasing which completeness decreases. V-measure, in this case, is maximal approximately at the point in which the two cross.

In figure 5 we are able to compare results across the two clustering algorithms – k-means and hierarchical – as well as across data sets of different sizes and class distributions. We can observe similar trends in the behavior of the clustering algorithms across data sets. K-means tend to achieve an optimal clustering with fewer clusters than the agglomerative clustering approach. Moreover, on the document clustering task, this optimal approach is considerably higher than the maximum yielded by the agglomerative approach. This allows us to conclude that with the described features k-means is better suited to these tasks. While, neither shows overwhelming success – these are naive feature spaces and algorithms – we can see that document clustering is a considerably easier clustering task than pitch accent type clustering, despite the larger feature space, and wider class distribution.

6 Conclusion

We have presented a new external cluster evaluation metric, V-measure. We have empirically demonstrated V-measure's satisfaction of some formal desirability criteria, as well as it's ability to evaluate document and pitch accent clustering solutions with respect to the criteria of homogeneity and completeness.

We believe that validity addresses some of the problems that affect other cluster measures. 1) It evaluates a clustering solution independent of the clustering algorithm, size of the data set, number of classes and number of clusters. 2) It does not require its user to map each cluster to a class. Therefore, it only evaluates the quality of the clustering, not a post-hoc class-cluster mapping. 3) It evaluates the clustering of every data point, avoiding the "problem of matching". 4) By evaluating the criteria of both homogeneity and completeness, validity is more comprehensive than those that evaluate only one. 5) Moreover, by evaluating these criteria separately and explicitly, V-measure can serve as an elegant diagnositic tool providing greater insight into clustering behavior.

Acknowledgments

The authors thank Martin Jansche, Sasha Blair-Goldensohn and Kapil Thadani for their insights and

feedback. This work was supported by DARPA under GALE grant number blah.

7 References