Published in final edited form as:
Electron Commer Res Appl. 2012 ; 11(2): 152–158. doi:10.1016/j.elerap.2011.12.006.

RSQRT: AN HEURISTIC FOR ESTIMATING THE NUMBER OF CLUSTERS TO REPORT

John Carlis and
Computer Science and Engineering; University of Minnesota, 4-192 Keller Hall; 200 Union St. SE, Minneapolis, MN USA 55455, Tel: 612-625-6092; Fax: 612-625-0572

Kelsey Bruso
Unisys Corporation, 2470 Highcrest Rd, Roseville, MN USA 55113, Tel: 651-635-7621

John Carlis: carlis@umn.edu

Abstract

Clustering can be a valuable tool for analyzing large datasets, such as in e-commerce applications. Anyone who clusters must choose how many item clusters, K, to report. Unfortunately, one must guess at K or some related parameter. Elsewhere we introduced a strongly-supported heuristic, RSQRT, which predicts K as a function of the attribute or item count, depending on attribute scales. We conducted a second analysis where we sought confirmation of the heuristic, analyzing data sets from theUCImachine learning benchmark repository. For the 25 studies where sufficient detail was available, we again found strong support. Also, in a side-by-side comparison of 28 studies, RSQRT best-predicted K and the Bayesian information criterion (BIC) predicted K are the same. RSQRT has a lower cost of $O(\log \log n)$ versus $O(n^2)$ for BIC, and is more widely applicable. Using RSQRT prospectively could be much better than merely guessing.

Keywords

Bayesian information criterion; clustering; data analytics; e-commerce; heuristic; spiral visualization

1. INTRODUCTION

Clustering can be a valuable tool for analyzing large amounts of data such as business transactions, network events, customer profiles, and activity levels. It is commonly used in e-commerce, data mining, or science (Lin and Dyer 2010).Among other issues, anyone who clusters faces a vexing problem: how many item clusters, K, to report. Ideally some algorithm decides the number of clusters. Unfortunately, as several authors lament, the clustering literature offers little help in predicting K. For example, Dubes (1987) says that "one of the most venerable problems in cluster analysis is: how many clusters are in the data?" Cadez and Smyth (1999) refer to the :"ever-thorny question of how many clusters are being suggested by the data." And Jain and Moreau (1987) indicate that a "very difficult problem in cluster analysis is to determine the number of clusters present in a data set."

Correspondence to: John Carlis, carlis@umn.edu.

Therefore, one must guess at *K* or some other related parameter with each of the three computational frameworks available in the clustering literature: Euclidean, model-based and complexity theory.

In the Euclidean framework cluster analysis techniques compare items using distance measures (Euclidean, Mahalanobis, Jaccard, etc.) (Restle 1959, Everitt 1993, Gordon 1981, Jain and Dubes 1988, Hartigan 1975, Tversky 1977). Euclidean technique authors are largely silent about predicting *K*, focusing instead on time complexity, statistical soundness, evaluating goodness given a *K* value, etc. So one must resort to trying every possible *K* value (which is often computationally infeasible), or to guessing. For example, a researcher might: for each *K*-means analysis, guess what *K* to supply as an input parameter; after each hierarchical clustering, guess where to cut a dendrogram to determine *K*; and, for either technique, guess which of several, often-conflicting goodness measures yields the "best" *K* (Jain and Dubes 1988, Hartigan 1975, Kaufman and Rousseeuw 1990, Dunn 1974, Gyenesei 2000, Milligan et al. 1983, Halkidi et al. 2000, Turenne 2000).

In the model-based framework, one hypothesizes a mixture of underlying probability distributions generating the data with each mixture component representing a different cluster (Dubes 1987, Fraley and Raferty 1998). With Bayesian theory, one can sometimes predict the number of components, but "even for the relatively simple Gaussian mixture model, the posterior cannot be calculated in closed form and must either be approximated analytically or estimated via sampling" (Smyth 2000). The Bayesian information criterion (BIC) can approximate the Bayesian model, but for only about a half-dozen components, beyond which it poses a substantial computational burden(Fraley and Raferty, 1998, 2000; Banfield and Raferty 1993). More importantly, one must still guess which model (spherical, diagonal, non-diagonal, etc.) best represents the cluster shapes (Bensmail and Celeux 1996).

In the complexity theory framework, global structure emerges from local activity rules; a system of elements (items) settles into *K* attractors (clusters); and power laws predict the element distribution(Lewin 1992).Unfortunately, complexity theory authors too are largely silent about *K*. A notable exception is Kauffman (1993) who observed a roughly square root relationship between elements and attractors, for example, in a Boolean network of 100 elements there were eight attractors, and one with 1,000 elements had 33 attractors. Others have observed organisms where the number of cell types was roughly the square root of the number of genes (Lewin 1992).

Whatever the framework, generally one must guess repeatedly, somehow picking an initial *K*, clustering, examining the results, picking a hopefully-better *K*, re-clustering – and so on until one homes in on a value of *K* that yields a satisfactory story. Of course, the analysis may not be so simple. One might try several *K* values at once, experiment with distance measures, analyze subsets of the data, etc. Sometimes a less pejorative term than "guess repeatedly" for this process is used, such as calling *K* "empirically chosen"(Bertone 2001)– but one still must guess.

Intrigued by Kauffman's square root observation, we set out to determine if it fit the *K* values that researchers actually have reported, and, more generally, to discover patterns in those *K* values that could assist in choosing *K*. To that end, we performed two analyses of reported *K* values, the first reported in Carlis and Bruso (2010), developed an heuristic predictive of *K*, and found that that the heuristic is a possible substitute for BIC, since it is cheaper and more widely applicable than BIC.

In the following sections we describe the RSQRT heuristic, and the data and results from the two studies. We then compare RSQRT with BIC, and related work. We finish with a discussion about aspects of RSQRT and the possibility of using it prospectively.

2. THE RSQRT HEURISTIC

Thus motivated, we began acquiring clustering articles, keeping track of a study's K reported clusters, and n items, where an item is a set of variable values of interest to a user. An item might be a purchase, a network packet delivery, a customer's demographic characteristics, or a protein's expression levels under differing experimental conditions. Since it quickly became apparent that Kauffman's notion would not be sufficient, we also tracked each study's t attributes, and each attribute's *scale*, which, broadly, is nominal (name, category, true/false, color, etc.) or numeric (temperature, weight, percentage, count, distance, etc.). Think of each study's data represented in a simple table with items as rows, attributes as columns, and result clusters as groups of rows with each row assigned by the clustering process to exactly one cluster.

After examining a few studies we devised a *recursive square rootheuristic*, RSQRT, extending Kauffman's notion two ways. First, we noticed that sometimes the value of $K_{reported}$ (always an integer) was closer to $t^{0.5}$ than to $n^{0.5}$, depending on the scale of the variables. So RSQRT has a count-of-things parameter, which is n if a study's attributes are all numeric, and is t if they are all nominal or of mixed scales. We note that nominal scales seem to dominate for clustering.

Second, while the square root of n or t is often far from the $K_{reported}$, we noticed that the fourth root, or higher, is often much closer to $K_{reported}$. So RSQRT produces a list of predicted values for K following this formula for x things, where x is n or t, as follows:

$$\text{RSQRT}(x)=\{x^{2\wedge(-r)};r=1,2,\ldots,r_{max};x^{2\wedge(-r_{max})}<2.25\}$$

and r is called the *recursion level*.

For example, a study with 10,000 things has RSQRT(10,000) = {100, 10, 3.1, 1.76}, and the maximum possible r is $r_{max} = 4$. RSQRT repeatedly takes square roots, terminating when its last value is below 2.25, because the next list value would be closer to 1 than 2, and $K = 1$ would add no value. If this example study's $K_{reported} = 3$ then its $K_{best\text{-}predicted} = 3.1$, its $r = 3$ (eighth root), and its difference, $d = | K_{reported} - K_{best\text{-}predicted} | = 0.1$.

The heuristic predicts that $K_{reported}$ will be close to a $K_{predicted}$ value, that is, d will be small. Do not misinterpret d, which means difference, not error. We do not presume to judge the goodness of the values of $K_{reported}$ that researchers and reviewers found worthy. Instead, we are looking to find helpful patterns in what they report.

3. METHODS

To assess the goodness of the heuristic's fit and the breadth of its applicability, we sought a broad collection of articles coming from mostly biomedical disciplines, with only 10% or so of the studies related to e-commerce. The studies employed many different clustering techniques. We focused on real datasets, and excluded synthetic data sets. A few articles had more than one "study," analyzing more than one dataset or using more than one clustering technique. Of course, we expect that most researchers try more than one technique, distance measure, or K, but only report the best results, or just those that offer insight to a problem, and were acceptable to reviewers – ones that help tell a coherent story.

We acquired 251 studies but dismissed 25 because, although the authors used clustering analysis techniques, they had predetermined K, and therefore, we think, classified, not clustered. We note that predetermined determined clusters might be called *natural clusters*,

but their use is post-clustering. This generally involves assigning a new item to one of those clusters, that is, classifying it, not clustering. The remaining 226 studies encompass 44 different subject areas (Table 1), and use several dozen different clustering techniques (Table 2), and use whatever distance measures were appropriate for their data. These studies come from various places, mostly journals, with about half found via PUBMED.

4. RESULTS

This milieu exhibits a “serial and periodic” characteristic, for which a spiral display can be valuable (Carlis and Konstan 1998). Therefore, with Figure 1's spiral we compactly present various study properties: scale, r, d and the $K_{reported}$ distribution. A spiral of Archimedes is augmented so that square and higher roots are readily discernible, for example, the ray passing from the origin through 256, and 16, the square root of 256, is where the next inner lap of the spiral touches the ray ($r = 1$); the fourth root (4) is on the second inner lap ($r = 2$), etc. For each of the 226 studies a symbol appears on the spiral showing its:

attribute *scale* kind by its fill (bluish = numeric; amber-ish = nominal or mixed);

t(amber shape) or n(blue shape) value by its position, with position numbers on the outside of the spiral line.

d category by its darkness (darker means a better fit with the heuristic), with these categories:

- darkest ($d \leq 0.5$; which rounds to an exact match);
- dark($0.5 \leq d < 1.0$);
- pale($1.0 \leq d < 2.0$);
- palest ($d \geq 2.0$).

r by $r+2$ sides to its shape (more sides means higher recursion), for example, $r = 1$ appears as a triangle.

Note that studies may be stacked at a position, for example, several studies stack at 30, where the law of large numbers begins to give a study statistical significance. Also on the spiral is the frequency distribution of $K_{reported}$, depicted with italicized numbers on the inside of the spiral line, for example, 24 studies have $K_{reported} = 5$.

Here is a brief description of one study for each kind of attribute scale, plus how it appears on the spiral.

All nominal. A group of $n = 136$ people, normal or diagnosed with schizophrenia, were measured on $t = 70$ variables. Using intra-group agreement scores calculated by hand, groups of individuals similar among themselves were analyzed to detect common response patterns. Zubin (1938) reports nine clusters of people. Applying RSQRT to t we get RSQRT(70) = {8.3, 2.8, 1.7} An arrow points to the medium dark amber ($d = 0.7$) triangle ($r = 1$) for this study, which is one of five studies reporting nine clusters.

All numeric. A group of $n = 655$ cancer and control patients were clustered using K-means according to their measured food-intake ($t = 1$). Chen et al. (2002) report six clusters of patients. Applying RSQRT to n we get RSQRT(655) = {25.59, 5.05, 2.24}. An arrow points to the medium dark blue ($d = 0.95$) square ($r = 2$) for this study, which is one of fifteen studies reporting six clusters.

Mixed. Electromicroscopy is used to measure $t = 16$ characteristics of $n = 42$ species of pollen. There are twelve continuous variables (numeric) measuring geometric properties of the trilobed grains and four discrete (nominal) variables measuring, by unreported

means, the presence or absence of patterns. Applying RSQRT to t, we get RSQRT(16) = {4}. Small et al. (1971) report four clusters of species. An arrow points to the darkest amber ($d = 0.0$) triangle ($r = 1$) for this study, which is one of forty studies reporting four clusters.

Next, our examination of the results focuses on r and d.

For r, the recursion level, Figure 1 and Table 3 show that fewer numeric scale studies had $r = 1$ (blue triangles) than $r > 1$ (blue squares, etc.), but the opposite holds for nominal or mixed studies (amber shapes). This is sensible since a study's n value generally is considerably larger than its t value. Only a few studies have a t large enough for the formula to calculate an eighth or sixteenth root, and none of those roots had the best d. So Kauffman's square root of n notion (blue triangle) fits just 58 studies, which is about one of three numeric studies and about one of four studies overall.

For d, the difference between predicted and reported, the predominantly dark (blue or amber) shapes in Figure 1 and the distribution of d values in Table 4 strongly support the heuristic. The median d is 0.4 and the average d is 1.3, both of which are much smaller than Kauffman's examples led us to expect. For over half of the studies d exactly matches the $K_{best\ predicted}$ (darkest shapes), that is, $d \leq 0.5$, and we round to the nearest integer to obtain $K_{predicted}$. Also, $d \leq 1.0$ for >3 out of four studies, and $d \leq 2.0$ for >9 out of ten studies. The distribution of d values is approximately the same for each of the attribute scales as it is overall. (Note the darkness pattern for blue and amber shapes in Figure 1.)

We looked in more detail at the worst fitting studies, with $d > 2$ (palest shapes), and found that all but two of the 21 such studies had one or more extenuating factors qualifying its poor fit. First, the four studies with the largest d values cluster data with more complex distances than simple numeric ones, using either multiple genomic sequence segments or molecule shapes. Excluding just these four studies lowers the average d to a remarkable 0.8. Second, eight large d studies come from articles where other studies analyzing its dataset had an exactly matching $K_{reported}$. Third, three large d studies, although they used real data, came from a software vendor's tutorial intended to illustrate clustering techniques rather than report refereed scientific results. Fourth, two large d studies used expert opinion, one employing a self-described "poor man's" clustering technique, while the other was based on a now-obsolete soils taxonomy. Finally, seven large d studies used *hierarchical agglomerative clustering* (HAC), which yields a dendrogram not clusters, and researchers had to guess about where to cut the dendrogram to form clusters. (Even including these poorly fitting studies, the median d for studies using HAC overall was 0.4.) Therefore, the RSQRT heuristic is even more strongly supported than it appears at first.

We also analyzed three other d properties, finding that each supports the heuristic. First, we verified that the recursive square root was a better fit than other functions, for example, cube root or log. Figure 2 shows the distribution of the count of studies versus the root for $K_{reported}$. These values were calculated using $\log(K_{reported}) \div \log(n$ or t, as appropriate). Figure 2 has large peaks at 0.5 (square root) and 0.25 (fourth root); and has minor peaks at 0.12 (eighth root) and 0.06 (sixteenth root). Further, there are no noticeable peaks at .33 or elsewhere, which, if present, would provide support for the cube root or other roots. Figure 3 plots d versus the reported K for log and for RSQRT. Note that for $K = 4$ and beyond, the d using the log is always larger than the d using RSQRT. There is no recursive application of log for this set of studies because $\log(\log(m))$ or $\log(\log(t))$ nearly always produces a value less than 1, which would imply no clustering. Second, the sign of $K_{reported} - K_{best\ predicted}$ is about evenly divided between positive (122 studies) and negative or zero (104 studies). Third, the paired t test shows RSQRT is statistically better ($p < 0.05$) than its opposite, that is, using t as an argument to RSQRT with numeric scale data, and n with nominal or mixed.

Finally, as shown in Figure 4, removing the $K_{reported} = 2$ or 3 studies, where d cannot be more than 1.0, somewhat worsens the d statistics, but does not qualitatively change our assessment.

Next, to validate the heuristic further, we conducted a second analysis, applying the heuristic against the University of California Irvine Machine Learning Repository, using the Fiebrink subset, where 25 studies met the criteria for our analysis of the heuristic(Fiebrink et al. 2005). The others lacked sufficient detail, clustered on sequences, or performed classification, not clustering.

Figure 5 shows the cumulative frequency distribution of the d values. It shows that in the second analysis most of the d values are small, and if added to Figure 1, would have dark values. Figure 4 also plots the distribution if small K values are excluded, ones where d cannot be large.

As shown in Figure 6, the exponent of $K_{reported}$ values peaks at .5 and .25 (as in Figure 2), and 92% (23 of 25) of the d values are less than 0.5 (which is exactly matching). Figure 7 shows a side-by-side comparison of the Machine Learning Repository data sets and the first analysis' data sets. Both collections show a local maxima near 0.5 (square root) and near 0.25 (fourth root). The original data set also shows small peaks near 0.12 (eighth root) and 0.06 (sixteenth root).

In sum, the two analyses each show strong support for the RSQRT heuristic. The k distribution, r distribution, d properties, and other characteristics are substantially the same in both.

5. RSQRT VERSUS BIC

BIC is a widely used technique for estimating the number of clusters in a data set.(Banfield and Raferty 1993; Bensmail and Celeux 1996; Fraley and Raferty 1998, 2000).In order to support our claim for the usefulness of RSQRT, we performed an experiment to compare, for the same data sets, the RSQRT best predictions against the reported BIC prediction, where the authors had to guess, perhaps repeatedly, which model best represents the cluster shapes.

We chose 27 journal articles that used BIC as a data analysis technique, from IEEE, PubMed, and Citeseer. These articles were all published in refereed journals and so had their BIC estimates and results vetted by peer review. We vetted the articles using the same criteria we used to subset our more general set of 251 journal articles. We rejected two because we could not determine n, the original number of items in the data set being analyzed; we rejected four because the studies used only synthetic data; we rejected ten because they used BIC for classification not for clustering; and we rejected one because it clustered sequences of data, which we know RSQRT does not handle well. This left 12 journal articles, three of which were part of the first analysis. They describe 28 studies.

We performed pairwise statistical analysis on the K values using the null hypothesis that the estimates generated by BIC and RSQRT. They are equal and the alternate hypothesis that the estimates are different. The t-statistic is 1.548. The two-tailed t critical value for 27 d.f. with a .05 significance level (90%) is 2.052. Since 1.548 is less than 2.052, we cannot reject the null hypothesis, meaning that we accept the null hypothesis that the BIC estimate and the RSQRT estimate are equal.

We used a paired statistic for analyzing whether the predicted value from BIC and the predicted value from RSQRT are the same. For the statistical analysis, our null hypothesis is

that the mean d value for the paired difference is 0: that is, the two predicted values are equal. The alternate hypothesis is that they are not equal. There were these statistics: BIC - RSQRT = 0; count = 28; average difference = 1.14; std. dev. = 3.897; t score = 1.548. The standard t score α = .050 for a two-tailed test with 27 degrees of freedom is 2.052. Since 1.548 is less than 2.052, we do not reject the null hypothesis. The p-value for this test is 0.1212. Since the p-value is greater than the α value of .050, again we do not reject the null hypothesis. In summary, we cannot reject the null hypothesis.

The consequences of this comparison are threefold. First, the analysis shows that from a statistical viewpoint, BIC and RSQRT estimate the same number of clusters in a data set. Soit appears that either technique may be used, although more confirming evidence is desirable.

Second, RSQRT, with a computational complexity of $O(\log \log n)$, has an advantage over BIC with a computational complexity of $O(n^2)$. To use BIC effectively, a researcher must guess at shapes, and calculate BIC across multiple K values against multiple statistical models to get a $BIC_{predicted}$ value for K. The effective computational cost is w choices for Ky model choices times $O(n^2)$. From a researcher's standpoint, BIC is burdensome.

Third, RSQRT applies more broadly than BIC. RSQRT applies to both nominal scale attributes and numeric scale attributes. With BIC you must pay close attention to nominal scale attributes to avoid having them dominate the numeric scale attributes.

6. OTHER RELATED WORK

Two other research efforts deserve mention here. Tibshirani et al. (2001) proposed the gap statistic procedure to generate a set of possible values for k and then identify the most likely K value. This is the value of K with the minimum value for the authors' gap statistic. The processing expense is perhaps equal to or perhaps slightly less than using BIC and the authors don't make a comparison directly against the BIC. However, as we suggest with RSQRT, Tibshirani et al. (2001) suggested that researchers must apply the gap statistic thoughtfully: "The results for the gap statistic are shown ... The estimated number of clusters is two … However, the gap function starts to rise again after six clusters, suggesting that there are two well-separated clusters and more less separated ones." This implies that K=2 gives the best value for the gap statistic, but K=6 might provide better support for the researchers' story.

Ben-David et al. (2007) addressed a related question about the goodness of the chosen K more than choosing K itself. They stated that "distressingly little is known about the theoretical properties of clustering. In particular, two central issues, the problem of assessing the meaningfulness of a certain cluster structure found in the data set and the problem of choosing K… which best fits the given data set are basically unsolved." Their paper described an approach to analyze cluster stability a means to address the first issue without directly addressing the second issue.

7. CONCLUSION

We presented here strong evidence that our heuristic successfully predicts the K values that researchers actually report with simple nominal, numeric, or mixed scale attributes, and does so across many subject areas and both manual and algorithmic clustering techniques. Certainly our heuristic fits more broadly than Kauffman's notion. However, questions remain: What are its limits? Why does it work? And is it useful?

We already know that our heuristic's limitations include predicting poorly for some complex numeric scales. We also expect it to not predict well with document clustering, for example, the Reuters-21578 at UCI, where an item by variable matrix formulation does not fit. We also suspect that it will predict poorly where subsets of the items are sequentially interrelated, because that makes it problematic to base calculations on the number of items, n. The heuristic might apply in areas other than clustering where picking K is an issue, such as matrix factorization (Berry et al. 2007).

Although, of course, we cannot say definitively why the heuristic fits so well with the K's that other researchers have reported, we can address several smaller questions about why what we discovered is sensible.

First, why n or t? With numeric scale attributes, clustering techniques assess more or less continuous differences, but with nominal ones they assess discrete matches. Considering a simple case. If numeric scale items A and B, and B and C are pair-wise near each other, then A and C will also be near – at worst, and uncommonly, the sum of the other pairs' distances. However, the similarity of two nominal scale items depends on how many attributes match (there is no "close" on one attribute), and if A and B match on half of the attributes and B and C match on half of the attributes, A and C might not match on any attribute. It may help to visualize the difference for a small (a t of three) problem. Consider, then a unit cube, for true/false nominal scale data, each item is a point on a cube corner, and two items are either in the same place or are far from each other – in different corners and there are only a few corners, so K cannot be large. Likewise, with mixed scale variables, even a few nominal scale variables makes the items appear near corners. In contrast, for numeric scale data normalized to a zero to one range (and so still depicted with a unit cube) each item is a point somewhere on the inside of the cube. Here it is unlikely that most of the items appear at or near corners, and so t and thus the number of corners does not influence K. So we think it sensible that researchers act differently with different scales.

Second, why recursive? This is a matter of pragmatics. In a report, clustering is not the end, but a means to help researchers make a convincing story, and a K in the thousands is beyond interpretation. A researcher rarely reports more than a few dozen clusters, so an $r > 1$ for a large n or t makes sense, as does an $r < r_{max}$, since r_{max} often corresponds to a K with too coarse a clustering to support an interesting, insightful story.

Third, why square root? Well, we think this is sensible since the square root is the simplest nonlinear reducer, and, furthermore, we found that its nearest competitors, cube root and log, predict poorly. However, we expect that more complicated scales will likely require more complicated heuristics.

Fourth, and perhaps most intriguing, does K come from data or people? That is, can a good K be determined from the data independent of observers, as implied by the literature quotes asking how many clusters there inherently "are," or must people judge that a K tells a good story? While the substitutability, in some cases, of RSQRT for BIC argues for inherency, this remains an open question. It may be that this questionis a matter for philosophy or cognitive science, and that Lakoff's embodied mind categorical metaphor is relevant (Lakoff and Nuñez 2001)

Wethink that researchers decide upon a K that best supports their ability to tell a compelling story about a data set. They explore it, and do not conclude that the result of a single execution of any technique shows an inherent number of clusters. Techniques such as BIC, Rand's statistic, mixture models, global optimum search, and others give the researcher some insight into a likely number of clusters to report. In reading the articles underlying our work the most compelling ones use multiple values for K, analyze pros and cons for each K

value, searching for a *K* supporting a story. Whether the researcher creates this list of possible *K* values using multiple subsets (training sets) of the data, or using multiple techniques, or using RSQRT is really secondary to what the different values of *K* reveal about the data set. Real data sets are messy and can support multiple interpretations or multiple insightful stories about them.

Whatever the ultimate source of *K*, for now "what *K*?" is a practical question for every researcher who clusters to address. The heuristic is not prescriptive, that is, it does not say one must report an RSQRT *K* value. However, it might be used prospectively as a replacement for plain guessing and as a low-cost replacement, where appropriate, for guessing with BIC. If the dataset has simple variable scales, then to use the heuristic prospectively one would: assess whether to use *n* or *t*; either get the RSQRT list by calculating or by inspecting the spiral; pick one (or perhaps several) of the candidate *K* values (perhaps also trying plus and minus one from the target or targets) to use as an initial *K* target. One would work as before: clustering and examining the suitability of the results; and so on. Thus the heuristic can provide researchers who must pick *K* with a starting point considerably cheaper than BIC and considerably better than guessing.

Acknowledgments

This work was supported in part by the U.S. NIH grant 1R01DE017734.

References

Banfield J, Raferty A. Model-based Gaussian and non-Gaussian clustering. Biometrics. 1993; 49:803–821.

Ben-David S, Pal D, Simon H. Stability of k-means clustering. Proceedings of 20th Annual Conference on Learning TheorySan Diego, CAJune 13–15, 2007:20–34.Also published as Lecture Notes in Computer Science. 4539SpringerBerlin, Germany2007;

Bensmail G, Celeux G. Regularized Gaussian discriminant analysis through eigenvalue decomposition. Journal of the American Statistics Association. 1996; 91:1743–1748.

Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. Algorithms and applications for approximate non-negative matrix factorization. Computational Statistics and Data Analysis. 2007; 52(1):155–173.

Bertone P, Gerstein M. Integrative data mining: the new direction in bio informatics. IEEE Engineering in Medicine and Biology Magazine. 2001; 20(4):33–40. [PubMed: 11494767]

Cadez, I.; Smyth, P. Technical report 99–16. Department of Information and Computer Science, University of California; Irvine, CA: 1999. Probabilistic clustering using hierarchical models.

Carlis, J.; Bruso, K. How many clusters to report: a recursive heuristic. Proceedings of the 2010 International Conference of the IEEE Engineering in Medicine and Biology Society; Buenos Aires, AR. August 31-September 4, 2010; Washington, DC: IEEE Computer Society Press; 2010. p. 1069-1072.

Carlis, J.; Konstan, J. Interactive visualization of serial periodic data. IEEE User Interface and Software Technology Symposium; San Francisco, CA. 1998.

Chen H, Ward MW, Graubard BI, Heineman EF, Markin RM, Potischman NA, Russell RM, Weisenburger DD, Tucker KL. Dietary patterns and adenocarcinoma of the esophagus and distal stomach. American Journal of Clinical Nutrition. 2002; 75:137–144. [PubMed: 11756071]

Dubes RC. How many clusters are best? an experiment. Pattern Recognition. 1987; 20(6):645–663.

Dunn J. Well separated clusters and optimal fuzzy partitions. Cybernetics. 4:95–104.

Everitt, B. Cluster Analysis. 3. Halsted Press; New York, NY: 1993.

Fiebrink, R.; McKay, C.; Fujinaga, I. Combining D2K and JGAP for efficient feature weighting for classification tasks in music information retrieval. Proceedings of the Sixth International Conference on Music Information Retrieval; London, UK. September 11–15, 2005; 2005. p. 42-49.

Fraley C, Raferty A. How many clusters? which clustering method? answers via model-based cluster analysis. Computing Journal. 1998; 41(8):578–588.

Fraley, C.; Raferty, A. Technical report. Vol. 380. University of Washington; Seattle, WA: 2000. Model-based clustering, discriminant analysis, and density estimation.

Gordon, A. Classification. Chapman and Hall; New York, NY: 1981.

Gyenesei, A. Technical report. Vol. 368. Turku Centre for Computer Science; Turku, Finland: Oct. 2000 Fuzzy partitioning of quantitative attribute domains by a cluster goodness index.

Halkidi, M.; Vazirgiannis, M.; Batistakis, Y. Quality scheme assessment in the clustering process. Proceedings of the Fourth European Conference on the Principles and Practice of Knowledge Discovery in Databases; Lyon, France. September 13–16, 2000;

Hartigan, J. Clustering Algorithms. John Wiley and Sons; New York, NY: 1975.

Jain, A.; Dubes, R. Algorithms for Clustering Data. Prentice Hall; Englewood Cliffs, NJ: 1988.

Jain A, Moreau J. Bootstrap Techniques in Cluster Analysis. Pattern Recognition. 1987; 20(5):547–568.

Kauffman, S. The Origins of Order. Oxford University Press; New York, NY: 1993.

Kaufman, L.; Rousseeuw, P. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons; New York, NY: 1990.

Lakoff, G.; Nuñez, R. Where Mathematics Comes From. Basic Books; New York, NY: 2001.

Lewin, R. Complexity – Life at the Edge of Chaos. University of Chicago Press; Chicago, IL: 1992.

Lin, J.; Dyer, C. Data-Intensive Text Processing with Map Reduce. Morgan and Claypool Publishers; San Francisco, CA: 2010.

Milligan G, Soon S, Sokol L. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1983; 5(5):40–47. [PubMed: 21869081]

Restle F. A metric and an ordering on sets. Psychometrika. 1959; 24:207–220.

Small et al. xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx. Xxxxxxxxxxxxxxx, xx, x, 1971, xxx-xxx.

Smyth P. Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing. 2000; 9(63):63–72.

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society B. 2001; 63(Part 2):411–423.

Turenne, N. PhD thesis. Université Louis-Pasteur; Strasbourg, France: 2000. Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles.

Tversky A. Features of similarity. Psychological Review. 1977; 84(4):327–352.

Zubin JA. A technique for measuring like-mindedness. Journal of Abnormal Psychology Research. 1938; 33:508–516.

Highlights

We propose a heuristic RSQRT to estimate the number of clusters, K, in a data set.

RSQRT applies well to numeric and nominal scale data attributes.

Results correlate with BIC estimations for k and RSQRT has much a lower computational cost.

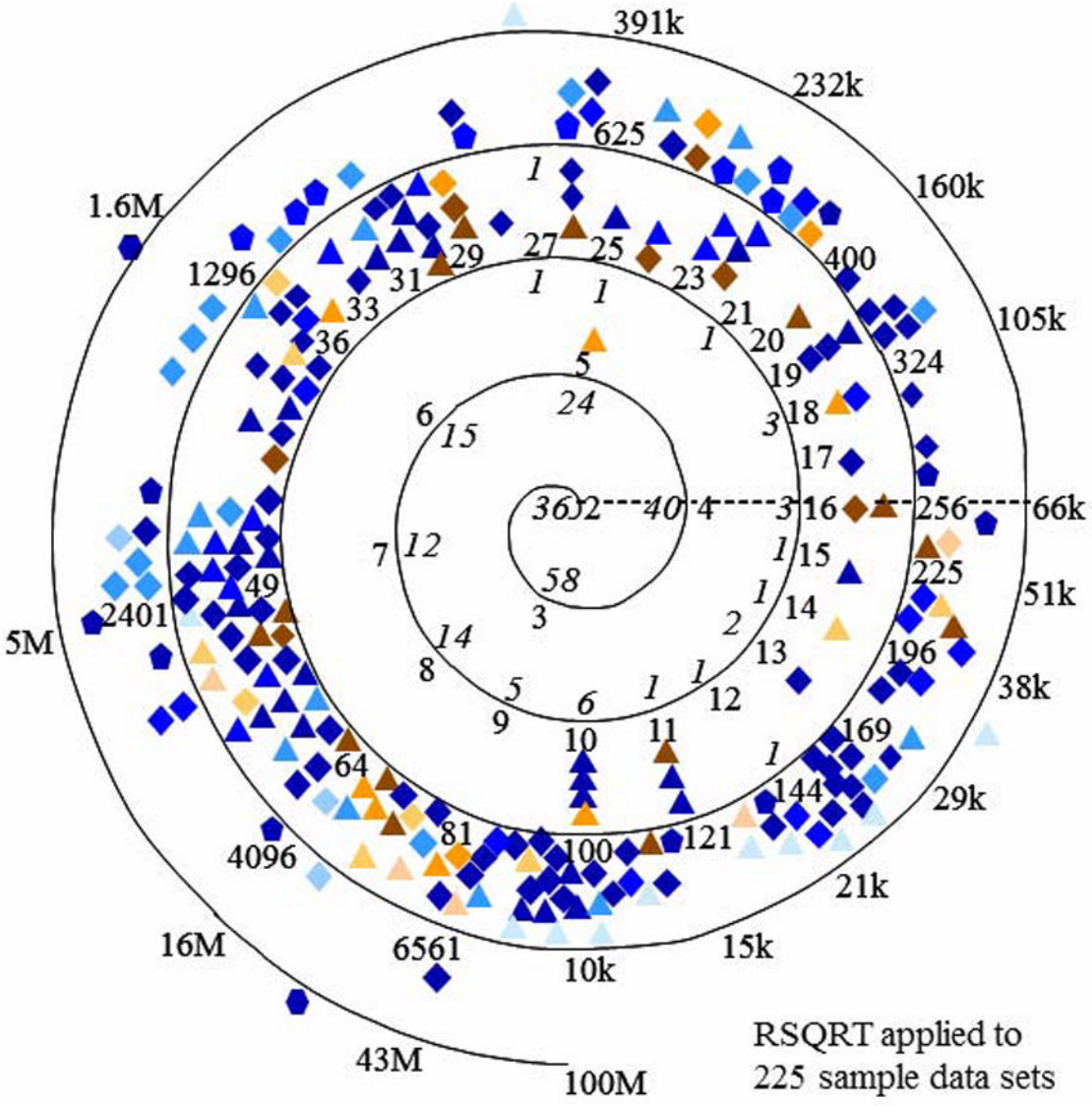

Figure 1.
A spiral display of RSQRT data

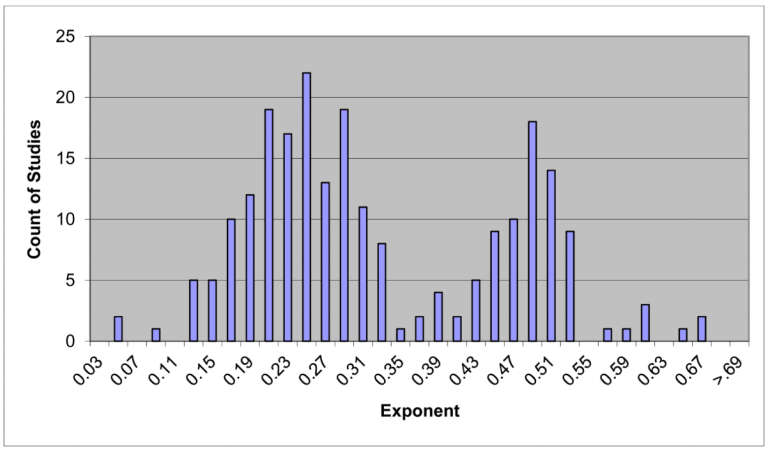

Figure 2.
Frequency distribution of the actual roots reported in our set of studies

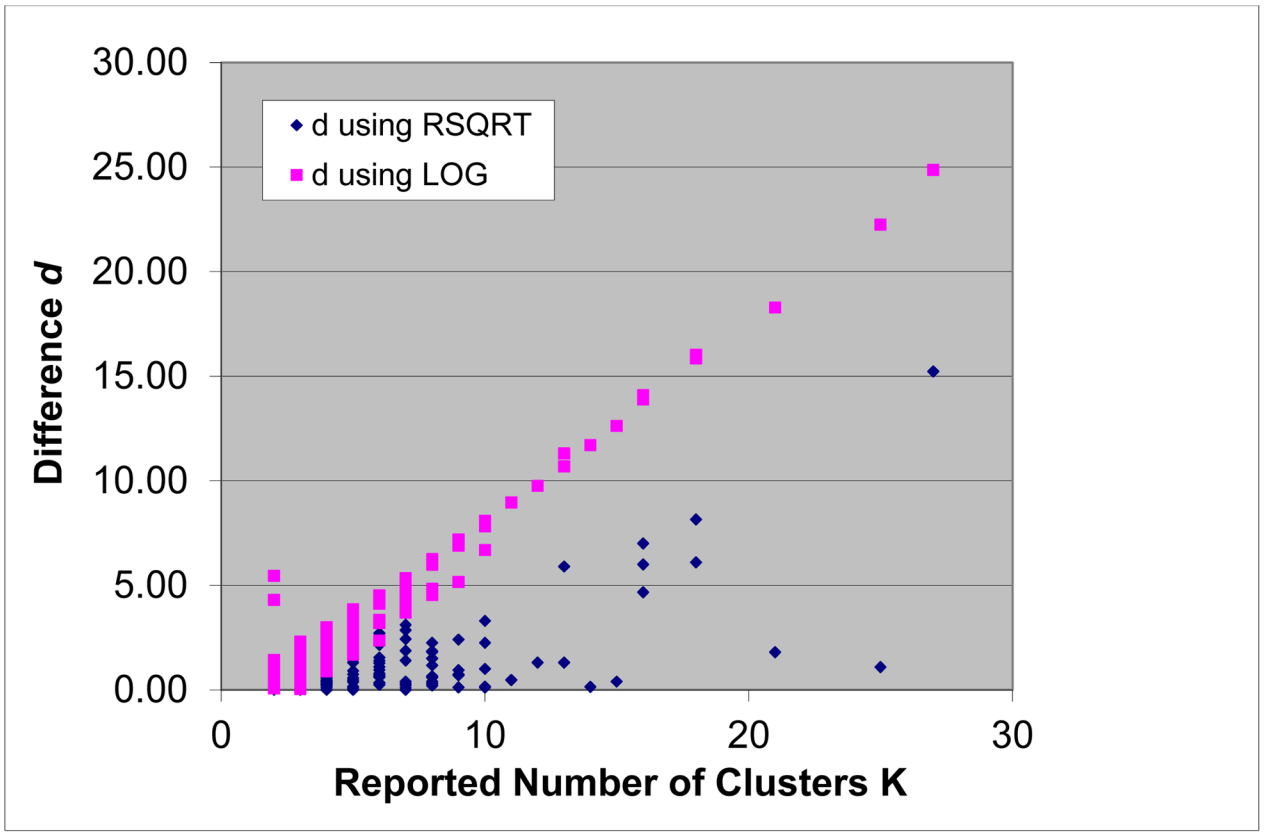

Figure 3.
Plots of *d* versus the reported *K* for log and for RSQRT

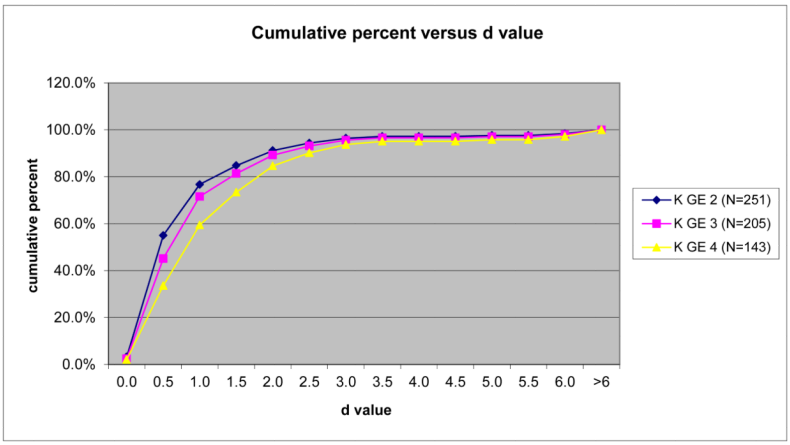

Figure 4.
Cumulative frequency distribution of *d* with differing minimal *K*

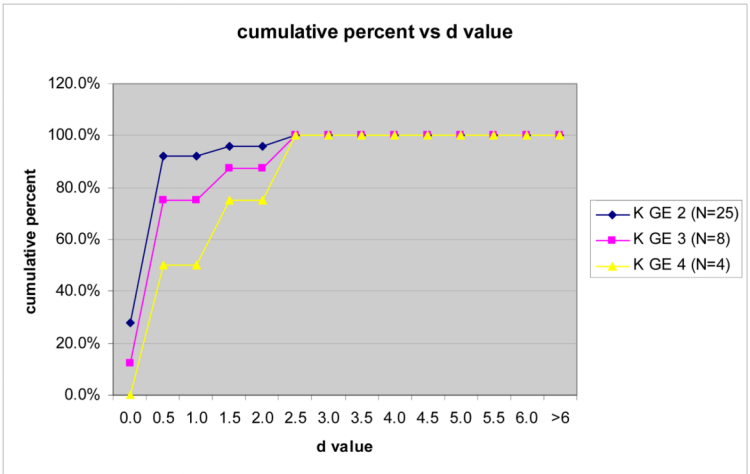

Figure 5.
Cumulative frequency distribution of *d* for Machine Learning Repository data sets

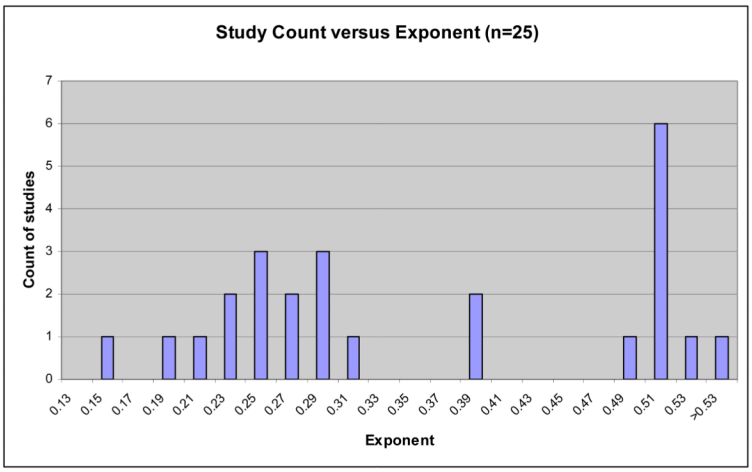

Figure 6.
Count of studies versus actual exponents of reported K versus the t or m values

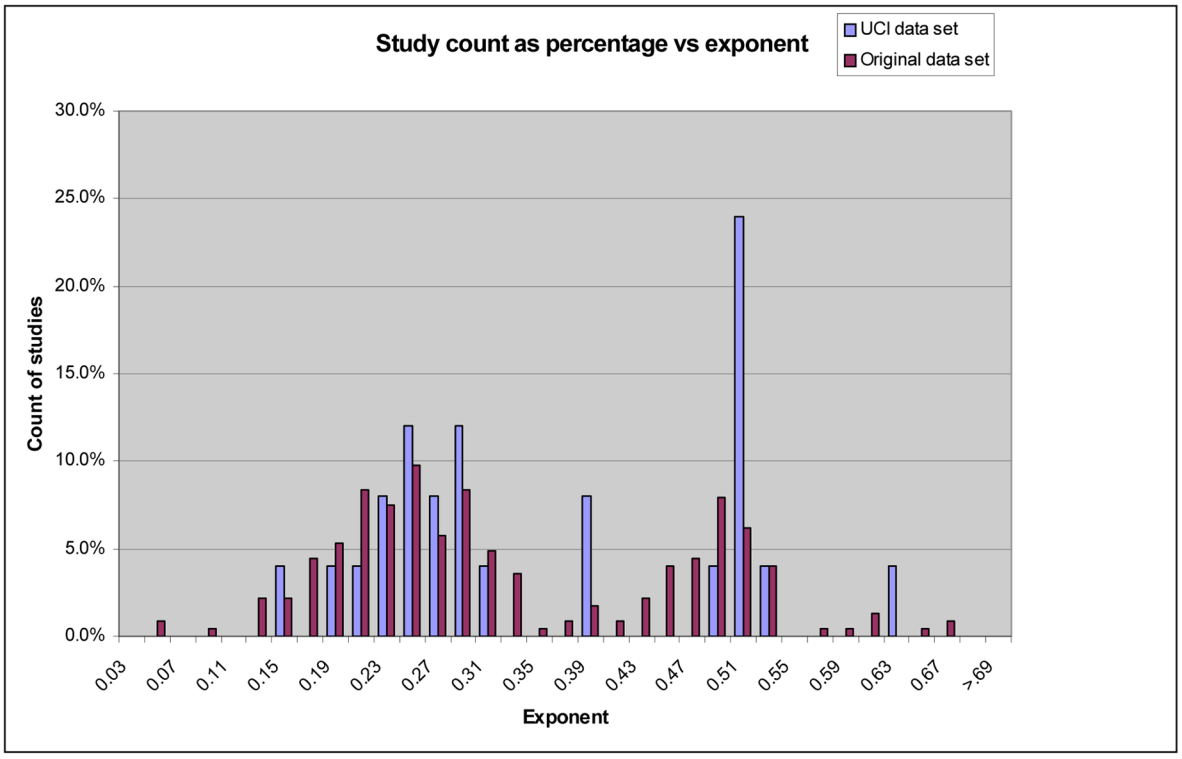

Figure 7.
Comparison of Machine Learning Repository data sets and first analysis data sets

Table 1

Frequency Distribution of Subject Areas

Area	Count
Anesthesiology	1
Animal science	1
Atmospheric science	1
Audiology	1
Brain injury	1
Geography	1
Nursing research	1
Oceanography	1
Psychooncology	1
Psychosomatic research	1
Acoustics	2
Adolescent psychology/child devel.	5
Agriculture	3
Alcohol abuse	7
Anthropology	5
Archaeology	20
Art	3
Biology/microbiology/neurobiology	15
Biomedical engineering	6
Business/economics	13
Chemistry/biochemistry/geochemistry	5
Radiology	1

Area	Count
Cluster analysis	15
Computer science – (web usage)	1
Dentistry / periodontology	3
Ecology	6
Education	3
Environmental systems	2
Epidemiology	3
Food science / nutrition	7
Gastroenterology	4
Genetics	8
Geology	7
Kinesiology/comparative/physiology	10
Market research/tourism	5

Area	Count
Medicine/pediatrics/gerontology	20
Neurology	1
Neurophysiology	2
Oenology	2
Pharmacology	2
Psychiatry	8
Psychology	18
Schizophrenia research	4
Total	**226**

Table 2

Frequency Distribution of Clustering Techniques

Clustering Technique	Count
Average-linkage	10
Between-group linkage	2
Complete-linkage	5
Cross-validated likelihood	2
Eigenvalue decomposition	2
Exact Bayesian inference	2
Finite mixture model	3
Hierarchical agglomerative clustering	28
Intersection	2
k-means	48
Single-linkage	13
Smith-Waterman	2
Sum of squares	3
UPGMA	7
Ward's	32
WPGMA	2
Other (14 techniques)	17
No technique credited	46
Total	**226**

Table 3

Distribution of study recursion level, *r*, of $K_{best\text{-}predicted}$ values by RSQRT parameter *n* or *t* and their total

***r* for $K_{best\ predicted}$**	***n;* blue**	***t;* amber**	***t* + *n***	**Figure 1 shape: *r* + 2 sides**
1	58	31	89	△
2	103	15	118	◇
3	17	0	17	⬠
4	2	0	2	⬡
Total	**180**	**46**	**226**	

Table 4

Study ranges, medians and distribution of *d* category properties

Ranges	*K*: 2 to 737 *n*: 10 to over 28 million *t*: 1 to 550			
Medians	*K*: 5 *n*: 97 *t*: 20			
Maximum *d* in category	0.5	1.0	2.0	>2
Figure 1 shape darkness	darkest	medium dark	medium pale	palest
n count	100	39	25	16
t count	22	10	9	5
n count + *t* count	122	49	34	21
Cumulative count	122	171	205	226
Category %	54	22	15	9
Cumulative %	54	76	91	100

$K_{reported}$	2	3	4	5	6	7	8	9	10	11	12	13	14	14	15	16
Frequency	36	58	40	24	15	12	14	5	6	1	1	2	1	1	1	3

17	18	19	20	21	22	23	24	25	26	27	plus a few others
0	3	0	0	1	0	0	0	1	0	1	