Documetric Forecast Algorithm

Dara O Shayda
dara@lossofgenerality.com

Nov 2nd 2013

Background

Let v_i be a value for time i e.g. stock value. Input finite list of tuples $(v_{i+1}, v_i, \text{vars}_v)$ to a machine learning algorithm e.g. SVM. 'vars_v' are any additional associate data e.g. high or low for stock v. The machine learning algorithms build a non-linear iterative universal approximator forecast such that:

$$\text{forecast}\,(v_i,\ \text{vars}_v) = v_{i+1}$$

In other words a forecast function. Note that vars_v is associate variables (as a function of i or time) for the parameter v.

Now add a new level of forecast-abilty i.e. documents related to the said tuples e.g. 10-K Sec filings of publicly traded companies or analyst reports.

The idea is to incorporate these documents into the trading via new forecasters.

These new forecasters have unstructured data i.e. mix of numeric Booleans (alarms) and text (parts of documents), hence the nam Documentric Forecast.

Algorithm

Start with a group of n documents f_n, each encoded into a TF-IDF vector. And to each document correspond a variable v_{j_i} e.g. stock price for company j at time i and each document to be its corresponding 10-K Sec filing section 7 titled "Management Discussion".

0. **Fix j**, this is the variable that the algorithm generates the forecast for
1. **Set a metric** d e.g. Euclidean Distance function (or any almost-metric e.g. Cosine Distance)
2. **Form clusters** of said vectors (documents) e.g. use MDS projection and visually extract the nearest-documents visually, or for example use K-means clustering
3. **Form tuples** $\left(v_{j_{i+1}}, v_{j_i},\ \text{cluster}_j,\ \text{vars}_v\right)$, where the cluster vector comprised of

$$\text{cluster}_j = \left(\frac{v_{1i}}{d(f_j,f_1)}, \frac{v_{2i}}{d(f_j,f_2)}, \ldots, \frac{v_{ni}}{d(f_j,f_n)} \right) \quad v_j \text{ and } f_j \text{ excluded}$$

4. **Choose a machine learning algorithm** and find forecast do include the cluster_j :

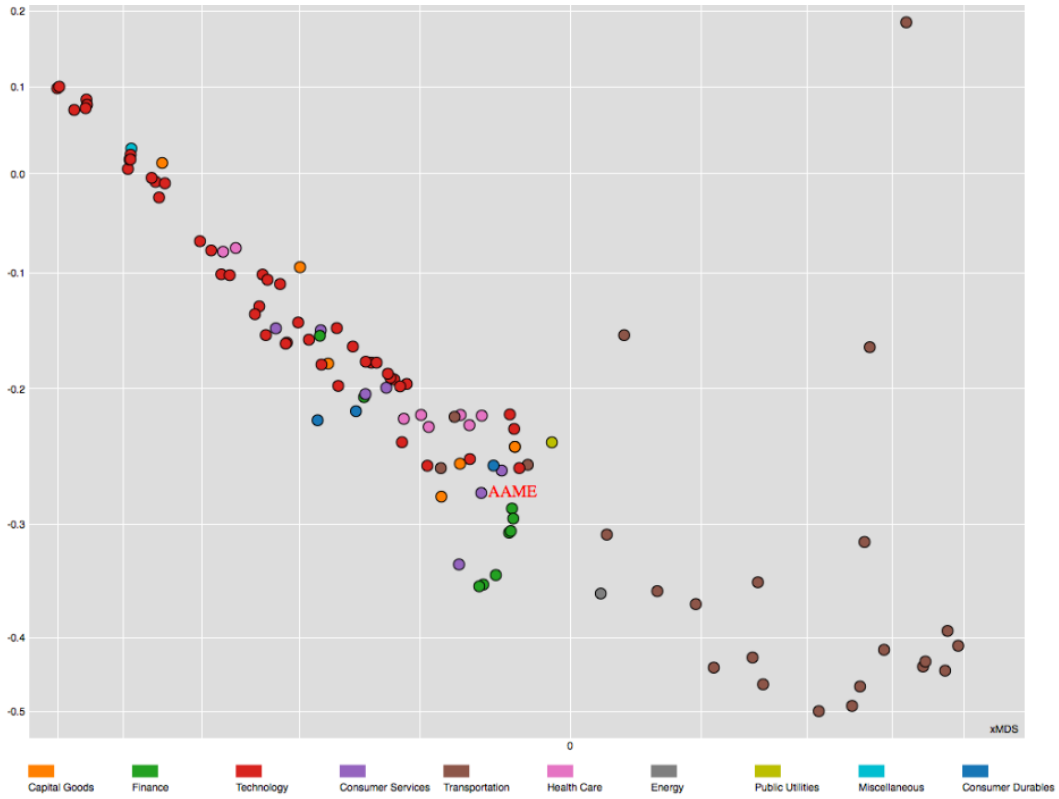
$$\text{forecast}\left(v_{j_i},\ \text{cluster}_j,\ \text{vars}_v\right) = v_{j_{i+1}}$$

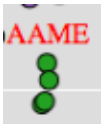
For more complex variations use a tuple like $\left(v_{j_{i+1}},\ v_{j_i},\ \text{cluster}_j,\ \text{vars}_v,\ \text{vars}_{\hat{v}}\right)$ where $\hat{v}$ is all the associate data (stocks high low ...) for all variables excluding v_j

5. **Self-Evaluate** to compute the accuracy of the forecast

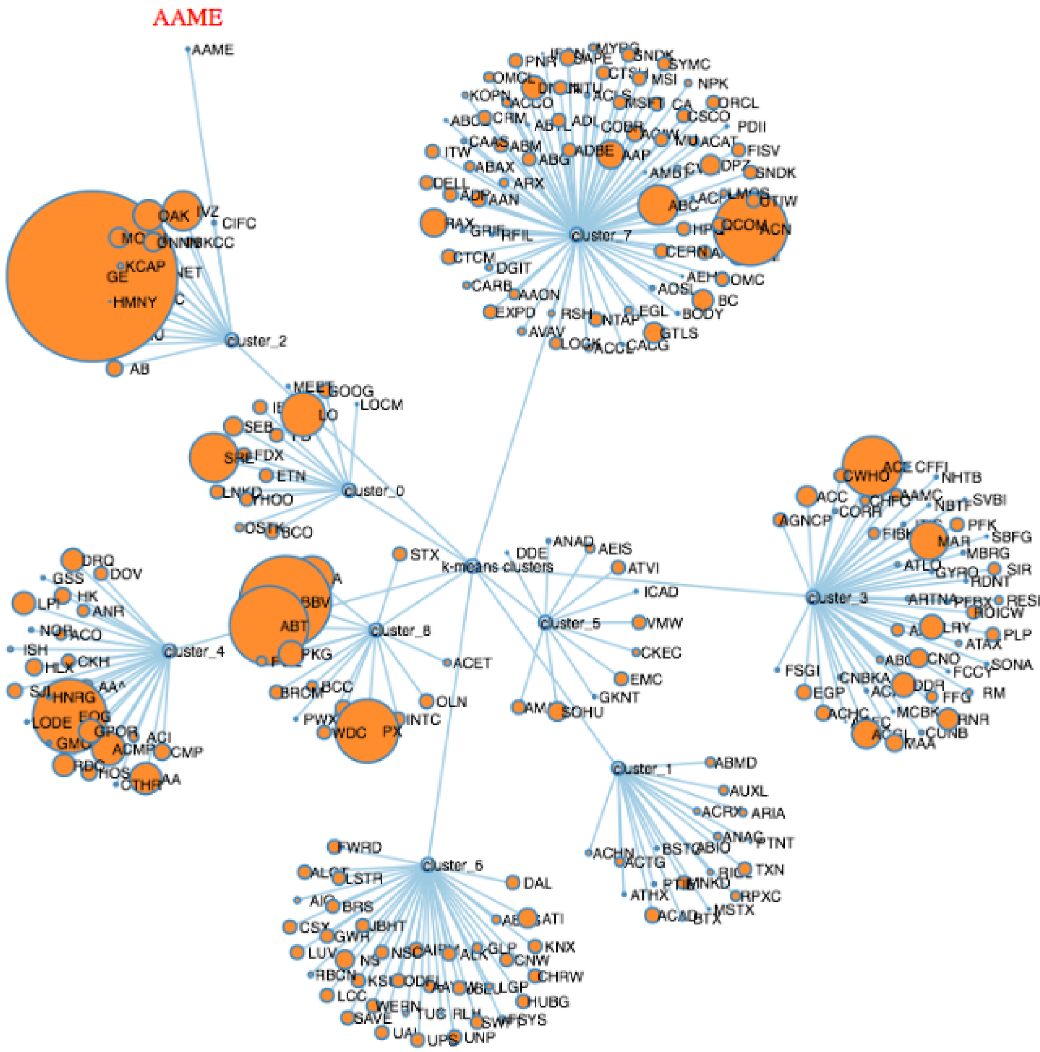
For #2 there are two or more ways of uncovering the subtle and mostly unknown clusters:

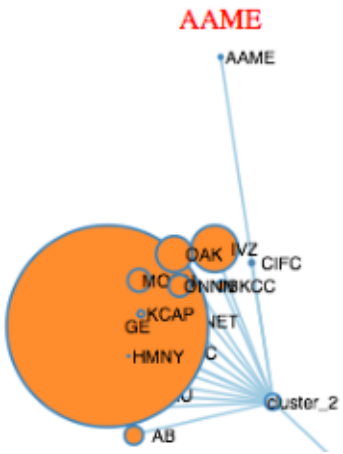
MDS projection





K-means clustering


AAME
AAME
k-means clusters
Cluster_0
Cluster_1
Cluster_2
Cluster_3
Cluster_4
Cluster_5
Cluster_6
Cluster_7
Cluster_8
CIFC
IVZ
OAK
HMNY
KCAP
AB
GOOG
LOCM
SEB
FDX
ETN
YHOO
STX
DDE
ANAD
AEIS
ATVI
ICAD
VMW
CKEC
EMC
GKNT
SOHU
ACET
OLN
INTC
WDC
PX
PKG
ABT
BBV
BCC
DRQ
DOV
GSS
HK
ANR
ISH
CKH
HNRG
GPOR
ACMP
CMP
ABMD
AUXL
ARIA
PTNT
TXN
RPXC
MSTX
ACTG
ATHX
ACHN
FWRD
DAL
LSTR
BRS
CSX
GWR
LUV
KNX
CNW
CHRW
HUBG
UNP
UPS
UAL
LCC
SAVE
RBCN
PNR
SYMC
NPK
ORCL
CSCO
PDII
FISV
SNDK
DELL
ITW
ABAX
ARX
ADBE
AAP
DGIT
CTCM
CARB
AAON
EXPD
AVAV
RSH
NTAP
EGL
BODY
GTLS
OMC
BC
AOSL
CERN
QCOM
ACN
HP
RFIL
ACC
CORR
AGNCP
NHTB
SVBI
PFK
SBFG
MBRG
SIR
RDNT
RESI
PLP
ATAX
SONA
FCCY
RM
RNR
CUNB
MAA
MCBK
EGP
FSGI
CNBKA
CNO
ATLO
ARTNA



Discussion

As distance between the documents f_j and f_k get larger, the contribution of v_k to the machine learning model becomes smaller via the quotient:

$$\frac{v_{ki}}{d(f_j,f_k)}$$

And as the distance decreases the contribution of v_k gets larger.

Some normalization for d metric function is necessary to avoid very large or very small numbers.

Choosing the sub clusters e.g. AAME example, increases the accuracy of the forecast and increases the useful contribution of the documents to forecasting.

Disclaimer

We provide a 'cloud' for computing the machine learning algorithms and data mining for forecast, classification and clustering applications. By no means we provide any guarantees on the models to be of certain accuracy or performance. It is the responsibility of the user to fine-tune the algorithms to produce the desired results. For that matter we provide a large variation for a certain algorithm and different versions of the algorithm.