CLAIMS

What is claimed is:

1. A computer-implemented method comprising:

providing a training data set comprising a set of data samples;

providing an auto-encoder, the auto-encoder encompassing a classifier;

performing an initial training of an embedding model of the auto-encoder, the embedding model being configured to map the set of data samples to a set of compressed feature vectors comprising feature elements, the set of compressed feature vectors defining a compressed feature matrix of the set of data samples;

providing a definition of a graph associated to the compressed feature matrix;

applying a clustering-algorithm to identify one or more node clusters of the graph;

applying a centrality algorithm to identify one or more central nodes of the one or more node clusters;

retrieving from an annotator one or more node labels for the one or more central nodes of the one or more node clusters, thereby creating annotated node labels;

propagating the annotated node labels of the one or more central nodes to other nodes of the graph, thereby creating propagated node labels; and

performing a further training of the auto-encoder, the further training comprising

training the embedding model of the auto-encoder and the classifier of the auto-encoder with the annotated and the propagated node labels, wherein the classifier is configured to predict the one or more node labels for the elements of the compressed feature vectors.

2. The method according to claim 1, the method further comprising:

repeating, at one or more iteration rounds, the applying of the clustering algorithm, the applying of the centrality algorithm, the propagating of the annotated labels and the performing of the further training of the auto-encoder until convergence.

3. The method according to claim 2, further comprising outputting the one or more node labels of the last iteration round before convergence as set of target items to be annotated.

4. The method according to claim 1, the method further comprising:

repeating retrieving from one or more annotators node labels for the one or more central nodes of the one or more node clusters.

5. The method according to claim 1, wherein the graph is defined by an adjacency matrix, the adjacency matrix being the product of the compressed feature matrix and a transpose of the compressed feature matrix.

6. The method according to claim 1, wherein applying the clustering algorithm comprises applying a graph-based clustering algorithm.

7. The method according to claim 1, wherein applying the centrality algorithm comprises applying a graph-based centrality algorithm.

8. The method according to claim 6, wherein the graph-based clustering algorithm is selected from the group consisting of:

k-spanning tree or minimum spanning tree algorithms;

shared nearest neighbor algorithms;

algorithms based on betweenness centrality; and

spectral clustering algorithms.

9. The method according to claim 7, wherein the graph-based centrality algorithm comprises:

approximating a product of a matrix exponential and a random probe vector of the adjacency matrix;

computing a diagonal of the adjacency matrix based on the product of the matrix exponential and the random probe vector; and

calculating node centralities based on the computed diagonal until a predefined number of the one or more central nodes has been detected.

10. The method according to claim 1, the method further comprising:

identifying one or more border nodes, the border nodes being defined as nodes which are situated at a border between two of the clusters.

11. The method according to claim 10, the method further comprising:

retrieving from one or more annotators node labels for the one or more border nodes.

12. The method according to claim 1, the method further comprising:

identifying one or more farthest nodes of the one or more clusters.

13. The method according to claim 12, the method further comprising:

retrieving from one or more annotators node labels for one or more farthest nodes of the one or more clusters.

14. The method according to claim 5, wherein the step of propagating the annotated node labels comprises applying a propagation function to the annotated node labels, wherein the propagation function is a function of the adjacency matrix.

15. The method as claimed in claim 14, wherein the propagation function is defined as a Chebyshev expansion $f(X)=\Sigma_{j=0} c_j P_j(x)$ where P_j is the j-th Chebyshev polynomial.

16. The method according to claim 1, further comprising:

training a cognitive model of a machine learning application using the training dataset.

17. A system comprising one or more processors for executing computer-readable instructions, the computer-readable instructions controlling the one or more processors to perform operations comprising:

providing a training data set comprising a set of data samples;

providing an auto-encoder, the auto-encoder encompassing a classifier;

performing an initial training of an embedding model of the auto-encoder, the embedding model being configured to map the set of data samples to a set of compressed feature vectors comprising feature elements, the set of compressed feature vectors defining a compressed feature matrix of the set of data samples;

providing a definition of a graph associated to the compressed feature matrix;

applying a clustering-algorithm to identify one or more node clusters of the graph;

applying a centrality algorithm to identify one or more central nodes of the one or more node clusters;

retrieving from an annotator one or more node labels for the one or more central nodes of the one or more node clusters, thereby creating annotated node labels;

propagating the annotated node labels of the one or more central nodes to other nodes of the graph, thereby creating propagated node labels; and

performing a further training of the auto-encoder, the further training comprising:

training the embedding model of the auto-encoder and the classifier of the auto-encoder with the annotated and the propagated node labels, wherein the classifier is configured to predict the one or more nodes labels for the elements of the compressed feature vectors.

18. The system according to claim 17 the system further comprising:

repeating, at one or more iteration rounds, the applying of the clustering algorithm, the applying of the centrality algorithm, the propagating of the annotated labels and the performing of the further training of the auto-encoder until convergence.

19. The system according to claim 17, the system further comprising:

outputting the one or more node labels of the last iteration round before convergence as set of target items to be annotated.

20. The system according to claim 18, the system being further configured to

repeat the step of retrieving from one or more annotators node labels for the one or more central nodes of the one or more clusters.

21. A computer program product comprising a computer-readable storage medium having program instructions embodied therewith, the program instructions executable by a processor to cause the processor to perform operations comprising:

providing a training data set comprising a set of data samples;

providing an auto-encoder, the auto-encoder encompassing a classifier;

performing an initial training of an embedding model of the auto-encoder, the embedding model being configured to map the set of data samples to a set of compressed feature vectors comprising feature elements, the set of compressed feature vectors defining a compressed feature matrix of the set of data samples;

providing a definition of a graph associated to the compressed feature matrix;

applying a clustering-algorithm to identify one or more node clusters of the graph;

applying a centrality algorithm to identify one or more central nodes of the one or more node clusters;

retrieving from an annotator one or more node labels for the one or more central nodes of the one or more node clusters, thereby creating annotated node labels;

propagating the annotated node labels of the one or more central nodes to other nodes of the graph, thereby creating propagated node labels; and

performing a further training of the auto-encoder, the further training comprising:

training the embedding model of the auto-encoder and the classifier of the auto-encoder with the annotated and the propagated node labels, wherein the classifier is configured to predict the one or more nodes labels for the elements of the compressed feature vectors.

22. The computer program product of claim 21, further comprising:

repeating, at one or more iteration rounds, the applying of the clustering algorithm, the applying of the centrality algorithm, the propagating of the annotated labels and the performing of the further training of the auto-encoder until convergence.