

CLAIMS

[Claim 1] A privacy protection based training method, wherein original data to be mined is generated from Internet-based activities and comprises m original samples, each original sample comprises a d -dimensional original vector x and an output tag value y , and m and d are natural numbers; and the method comprises:

generating, by a processor, n d -dimensional transform vectors π from the original data to be mined, each transform vector π being determined by a sum of yx of a plurality of original samples randomly selected from the m original samples, wherein the generating the n d -dimensional transform vectors π comprises one of:

in a first process, generating an m -dimensional vector σ , randomly determining $-v$ or v as a value of each dimension of σ , wherein the value of y is $-v$ or v , and v is a real number, and determining $\frac{1}{2}\sum_{i=1}^m(y_i + \sigma_i)x_i$ as a transform vector π , wherein y_i is an output tag value of an i -th original sample, x_i is an original vector of the i -th original sample, and σ_i is the i -th dimension of the vector σ ; and repeating the first process n times to obtain n transform vectors π ; or

in a second process, generating an m -dimensional vector w , randomly determining 0 or 1 as a value of each dimension of w , and determining $\sum_{i=1}^m w_i y_i x_i$ as a transform vector π , wherein w_i is an i -th dimension of the vector w , y_i is an output tag value of an i -th original sample, x_i is an original vector of the i -th original sample; and repeating the second process n times to obtain n transform vectors π ;

determining, by the processor, the n transform vectors π as training samples of a binary classification model; and

training, by the processor, the binary classification model based on the training

samples to obtain an outcome model for further data mining from the Internet-based activities.

[Claim 2] The method according to claim 1, wherein the binary classification model comprises: a Boosting algorithm, an SGD algorithm, an SVRG algorithm, or an Adagrad algorithm.

[Claim 3] A privacy protection based training device, wherein original data to be mined is generated from Internet-based activities and comprises m original samples, each original sample comprises a d -dimensional original vector x and an output tag value y , and m and d are natural numbers; and the device comprises:

a transform vector generating unit configured to generate n d -dimensional transform vectors π from the original data to be mined, wherein each transform vector π is determined by a sum of yx of a plurality of original samples randomly selected from the m original samples, wherein generating the n d -dimensional transform vectors π comprises one of:

in a first process, generating an m -dimensional vector σ , randomly determining $-v$ or v as a value of each dimension of σ , wherein the value of y is $-v$ or v , and v is a real number, and determining $\frac{1}{2}\sum_{i=1}^m(y_i + \sigma_i)x_i$ as a transform vector π , wherein y_i is an output tag value of an i -th original sample, x_i is an original vector of the i -th original sample, and σ_i is the i -th dimension of the vector σ ; and repeating the first process n times to obtain n transform vectors π ; or

in a second process, generating an m -dimensional vector w , randomly determining 0 or 1 as a value of each dimension of w , and determining $\sum_{i=1}^m w_i y_i x_i$ as a transform vector π , wherein w_i is an i -th dimension of the vector w , y_i is an output tag value of an i -th

original sample, x_i is an original vector of the i-th original sample; and repeating the second process n times to obtain n transform vectors π ; a training sample generating unit configured to determine the n transform vectors π as training samples of a binary classification model; and

a model training unit configured to train the binary classification model based on the training samples to obtain an outcome model for further data mining from the Internet-based activities.

[Claim 4] The device according to claim 3, wherein the binary classification model comprises: a Boosting algorithm, an SGD algorithm, an SVRG algorithm, or an Adagrad algorithm.

[Claim 5] A computer device, comprising: a memory and a processor, wherein the memory stores a computer program executable by the processor; and when the processor runs the computer program, the steps of any of claims 1-2 are executed.

[Claim 6] A computer-readable storage medium storing a computer program thereon, wherein when the computer program is run by a processor, the steps of any of claims 1-2 are executed.