TDK dolgozat

Marussy Kristóf
2014. október 22.

MŰEGYETEM 1782

Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Számítástudományi és Információelméleti Tanszék

Az intrinzikus dimenzionalitás átka a génkifejeződés adatok osztályozásában

TDK dolgozat

Készítette:
Marussy Kristóf
III. évfolyam

Konzulens:
Dr. Buza Krisztián
egyetemi adjunktus
Semmelweis Egyetem
Genomikai Medicina és Ritka Betegségek Intézete

2014. október 22.

MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Computer Science and Information Theory

The Curse of Intrinsic Dimensionality in Genome Expression Classification

TDK paper

Written by:
Kristóf Marussy
year III

Advisor:
Dr. Krisztián Buza
lecturer
Semmelweis University
Institute of Genomic Medicine and Rare Disorders

22nd October 2014

Contents

Rövid összefoglaló

A génkifejeződés profilok vizsgálata fontos eszköze az orvosi kockázatfelmérési, diagnosztikai és prognosztikai alkalmazásoknak (Alon et al., 1999; Bhattacharjee et al., 2001; Sotiriou et al., 2003). Az új generációs szekvenálási technológiák elterjedése a közelmúltban növekvő érdeklődéshez vezetett a génkifejeződés adatok prediktív osztályozása iránt.

Egy beteg génkifejeződés profilja több ezer gén kifejeződési értékét tartalmazhatja, ezért a génkifejeződési példányok sokdimenziós euklideszi tér vektoraiként ábrázolhatóak. Az ilyen nagy dimenziószámú terekben az osztályozóknak a „dimenzionalitás átka" néven ismert jelenségekkel kell megküzdeniük.

Az „átok" egyik legjelentősebb eleme a *csomósodás* (hubness), mely számos kutatás tárgyát képezte az utóbbi időben. A csomósodás a nagy *intrinzikus* dimenziójú adathalmazokban figyelhető meg *csomók* megjelenésének formájában (Radovanovic et al., 2010a). Csomók alatt olyan példányokat értünk, melyek meglepően sok más példányhoz hasonlítanak. Az adott alkalmazási területen a példányok a betegek génkifejeződési profiljait jelentik, hasonlóságuk távolságfüggvények, például az euklideszi távolságuk segítségével mérhető.

A csomósodás gyakran *rossz csomók* megjelenésével jár, melyek a hozzájuk hasonló példányoktól eltérő osztályba tartoznak. Az ilyen csomók csökkenthetik a tradicionális osztályozó algoritmusok pontosságát (Radovanovic et al., 2010a). A *csomósodás-alapú* (hubness-aware) osztályozókat úgy tervezték, hogy kihasználják a csomók jelenlétét, így elkerülhető a rossz csomók káros hatása (Tomasev et al., 2014; Tomasev és Mladenic, 2012).

Dolgozatomban összehasonlítom a leggyakrabban használt tradicionális és csomósodás-alapú osztályozók viselkedését a génkifejeződés adatokon. A kísérleteket nyilvános adatbázisokon (Alon et al., 1999; Bhattacharjee et al., 2001; Sotiriou et al., 2003), keresztvalidáció és statisztikai szignifikancia tesztek segítségével végzem.

Abstract

Gene expression profiles were found to be highly relevant for safety assessment, diagnostics and prognostics applications (Alon et al., 1999; Bhattacharjee et al., 2001; Sotiriou et al., 2003). Recent advancements in high-throughput sequencing technology lead to increasing interest in predictive classification models for gene expression data.

A patient's gene expression profile may contain expression values of thousands of genes. Therefore, gene expression instances are represented as vectors in very high-dimensional Euclidean space. Classification in such high-dimensional spaces is challenged by a collection of phenomena known as the curse of dimensionality.

One of the most prominent, recently explored aspect of the 'curse' is *hubness*, which was found to be related to the intrinsic dimensionality of the data (Radovanovic et al., 2010a). With hubness we mean the emergence of *hubs*, instances that are similar to a surprisingly large number of other instances. In our application, instances are patients' gene expression profiles and similarity may be determined by a distance function, e.g. Euclidean distance.

Hubs may frequently be *bad hubs*, which have a different class label than the instances they are similar to. These hubs may mislead traditional classification methods (Radovanovic et al., 2010a). However, *hubness-aware classifiers*, which were explicitly designed to take advantage of hubs, are able to mitigate hubness artifacts (Tomasev et al., 2014; Tomasev and Mladenic, 2012).

In my work, I explore the use of various traditional and hubness-aware classifiers on gene expression data. The comparisons are performed with cross-validation and significance testing on publicly available data sets (Alon et al., 1999; Bhattacharjee et al., 2001; Sotiriou et al., 2003).

1 Introduction

The *central dogma of life* describes the information flow inside cells between biopolymers (Lesk, 2012):

$$\text{DNA} \xrightarrow{\text{transciption}} \text{mRNA} \xrightarrow{\text{translation}} \text{Protein.}$$

The *deoxyribonucleic acid* (DNA) is a two-stranded double helix biopolymer consisting of nitrogen-containing nucleobases guanine (G), adenine (A), thymine (T), cytosine (C) and a phosphate-deoxyribose backbone. *Ribonucleic acid* has similar structure, however, it is single-stranded and contains uracyl (U) instead of thymine and ribose instead of deoxyribose.

Parts of the DNA sequence are *transcribed* into *messenger RNA* (mRNA) sequences in a process regulated by complex mechanisms according to changes in the environment inside and outside the cell. The mRNA transcriptome eventually finds its way to a *ribosome* of the cell, where it is *translated* into a polypeptide sequence built from the 21 (in some organisms, 23) *amino acids*. After various physical and chemical changes of the polypeptide chain, the end product is a *protein*.

Proteins are an extremely important class of biomolecules. They perform in virtually all kinds of functions within a cell: *structural* proteins are responsible for stiffness and rigidity, *enzymes* catalyse most intracelluar chemical reactions, *antibodies* bind to and neutralise foreign substances; proteins also regulate RNA transcription, signal information and transport substances between cells.

The information required to synthesise is contained in hereditary units of the DNA called *genes*. A gene is 'a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions' (Pearson, 2006). *Alleles* of gene are its variation found in nature.

The Human Genore Project has revealed that there are about 22 5000 genes in the human genome (International Human Genome Sequencing Consortium, 2004).

Gene expression studies look at the mRNA present in a specimen to determine what genes are currently used in the cell to synthesise proteins and what are their corresponding alleles. In *microarray* methods, a set of *probes* are created to capture alleles of interest (Antal et al., 2014). In contrast, RNA sequencing (RNASeq) methods, becoming increasingly popular doe to the widespread use of high-througput sequencing methods, can read the bases of the mRNA directly and hence discover yet unknown alleles.

Gene expression data may be used to categorise subtypes of diseases, effects of a treatment and find genes participating in specific disease a biochemical process (Kaminski and Friedman, 2002).

While gene expression measurement methods are gradually becoming more available, microarray and RNASeq measurements are still relatively expensive. For example, processing a sample with an Affymetrix® Gene Profiling Array widely used in clinical research, may cost \$400–\$700 including reagents and processing (Science Exchange, 2014).

The costs and the difficulty of recruiting a large number of patients for studies, especially in the case of rare diseases, means that gene expression studies are usually limited to a couple hundred specimens. This limitation of sample size combined with the large amount of genes and alleles makes data mining for gene expression data a 'small n, large p' problem (Aruliah et al., 2006). The large dimensionality of gene expression vectors ('large p') give rise to phenomena known as the ‘curse of dimensionality’ (Bellman, 1957), while the small number of vectors (‘small n’) means data sets are very sparse.

Machine learning and data mining applications remain to be challanged by the dimensionality and sparsity of gene expression data despite the importance of gene expression studies in clinical safety assessment, diagnostics and prognostics.

In this work, we explore the effects of *hubness*, a significant aspect of the ‘curse of dimensionality’, which is related to the *intrinsic dimensionality* data sets. After a discussion of related work in Chapter 2, we study how hubness appears in real-world gene expression data sets in Chapter 3. Chapter 4 concerns the effects of hubness in prediction task based on gene expression vectors. We conclude our study in Chapter 5.

2 Background

In this Chapter, we describe some data mining and machine learning approaches for gene expression.

2.1 Machine Learning for Gene Expressions

Machine learning tasks are usually categorised as either *unsupervised* or *supervised*.

In unsupervised tasks, only a set of *instances* $\mathscr{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ is available for the learning algorithm. In the domain of gene expression data mining, instances are vectors of gene expression values, i.e.

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,t}, \ldots, x_{i,p}) \in \mathbb{R}^p,$$

where $\mathbf{x}_{i,t}$ is the expression value of the ith patient and tth gene and p is the total number of genes considered. The vectors $\mathscr{X} = \{\mathbf{x}_i\}$ together form the data matrix X.

The unsupervised learning algorithm attemps to analyse the structure of the input data. For example, in *clustering* similar instances are grouped into clusters $\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_k$ by means of a partition $\mathscr{X} = \mathscr{C}_1 \amalg \mathscr{C}_2 \amalg \cdots \amalg \mathscr{C}_2$.

In supervised tasks, data *labels* are available in addition to the instances, thus, the data set takes the form $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ where y_i belongs to the set $\mathscr{Y}$ of all possible labels. In most problems, the learning algorithm attempts to construct a predictor M from $\mathscr{D}$. The predictor attempts to predict the label y^* of a yet-unseen instance $\mathbf{x}^*$. In *regression* tasks, the possible labels are the real numbers, i.e. $\mathscr{Y} = \mathbb{R}$. In *classification*, $\mathscr{Y}$ is a finite set of discrete *class labels*, for example $\mathscr{Y} = \{-1, +1\}$, where -1 means healthy tissue and $+1$ means cancerous tissue.

For a more comprehensive review, we refer to Antal et al. (2014, Chapter 8) in the context of gene expression data and Witten et al. (2011) in the context of general machine learning.

2.2 Dimensionality Reduction

Dimensionality reduction methods refer to method which can reduce the p-dimensional vectors $\mathscr{X} \subset \mathbb{R}^p$ into a more manageable form. The output is a set of lower-dimensional vectors $\mathscr{X}' \subset \mathbb{R}^q$ with $q \ll p$.

Because a cDNA microarray may be complementary to many thousands of genes, a large fraction of genes measured may be irrelevant to the disease or biochemical process studied. *Feature selection* methods attempt to select the relevant attributes, i.e. genes. This process is usually supervised, because we are interested in the genes that most likely determine the class labels.

On the other hand, *feature construction* methods create whole new representations of $\mathscr{X}$ not limited to the attributes already present. This projection can be done both in a supervised and unsupervised manner.

2.2.1 Feature Selection

Statistical indicators may be used to select relevant genes, including the t-statistic (Golub et al., 1999), twoing rule, information gain, gini index, max minority, sum minority and sum of variances (Murthy, 1998). An implementation of these rules is avaiable in RankGene (Su et al., 2003) software package.

Gene Ontology (Ashburner et al., 2000) is a dictionary (ontology) for expressing the roles of genes in biological processes. Under the null hypothesis, genes of a certain biological function are distributed among the overexpressed genes according to the hypergeometric distribution. Relevant genes may be selected according to the rejection of this null hypothesis.

Other hypothesis testing based methods for determining significance of genes include GSEA: Gene Set Enrichment Analysis (Subramanian et al., 2005) and SAM: Significance Analysis of Microarrays (Tusher et al., 2001).

2.2.2 Feature Construction

Principal Component Analysis (PCA) Principal Component Analysis is an unsupervised dimensionality reduction method that constructs and orthonormal system of *principal axes* $\{\mathbf{u}_1, \mathbf{u}_2, \dots \mathbf{u}_p\}$. Each principal axis corresponds to a linear combination of features (genes).

The principal axes are selected such that projection to the first q principal axes maximises the variance of projected data $\mathscr{X}'$. The projected vectors $\mathbf{x}'_i \in \mathscr{X}'$ are calculated from the corresponding original vectors $\mathbf{x}_i \in \mathscr{X}$ as follows:

$$\mathbf{x}'_i = (\mathbf{u}_1^T \mathbf{x}_i, \mathbf{u}_2^T \mathbf{x}_i, \dots, \mathbf{u}_p^T \mathbf{x}_i) \in \mathbb{R}^q. \tag{2.1}$$

Let

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{\mathbf{x} \in \mathscr{X}} \mathbf{x} \tag{2.2}$$

denote the data set mean and

$$S = \frac{1}{N} \sum_{\mathbf{x} \in \mathscr{X}} (\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{x} - \bar{\mathbf{x}}) \tag{2.3}$$

denote the data set covariance matrix. The principal axes are the eigenvectors $\{\mathbf{u}_i\}$ of S, sorted in decreasing order by the eigenvalues λ_i (Bishop, 2006, Chapter 12). Therefore, to obtain a q-dimensional representation of $\mathscr{X}$, we find the q larges eigenvalues of S and project the instance vectors to the corresponding eigenvectors.

Eigengenes The Eigengene method of gene expression feature construction (Alter et al., 2000) relies on Singular Value Decomposition (SVD) of the data matrix. The data matrix X is decomposed into a product of smaller matrices

$$X = U \Sigma V^T,$$

where Σ is a $q \times q$ orthogonal matrix, i.e. $\Sigma^T = \Sigma^{-1}$.

The left-singular vectors of X, which are the columns of U, correspond to q-many 'eigenarrays' and the right-singular vectors of X, which are the columns of V, correspond to q-many eigengenes. V contains the calculated expression levels of eigengenes in the instances of original data set $\mathscr{X}$, while U contains the expression levels of the genes of the original data set in the eigenarrays. The 'eigenexpression' matrix Σ connects the eigengenes and the eigenarrays.

2.3 Classification

In classification problems, the classifier M aims to maximize the classification *accuracy*

$$a(M) = P(M(\mathbf{x}^*) = y^*), \tag{2.4}$$

that is, the probabilty of correctly predicting a class label.

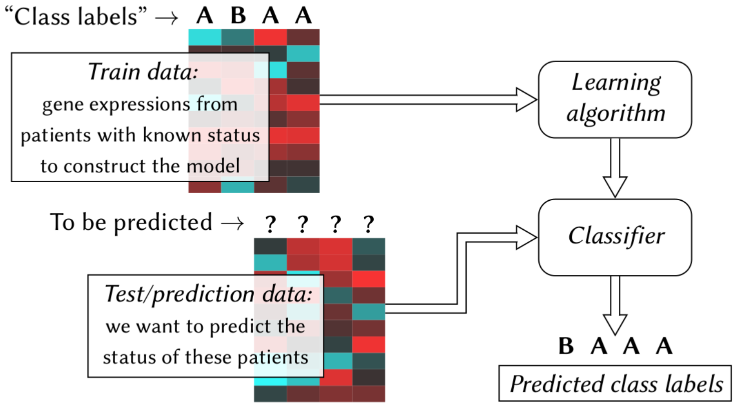

"Class labels" → A B A A
Train data:
gene expressions from patients with known status to construct the model
Learning algorithm
To be predicted → ? ? ? ?
Classifier
Test/prediction data:
we want to predict the status of these patients
B A A A
Predicted class labels

Figure 2.1 Classification for gene expression data..

To approximate the true accuracy $a(M)$, the labeled data set $\mathscr{D}$ is partitioned into disjoint a $\mathscr{D}_{\text{TRAIN}}$ training set and $\mathscr{D}_{\text{TEST}}$ testing set. The learning algorithm only has access to the training set when in construct the classifier M. The accuracy may be approximated as

$$\hat{a}(M) = \frac{|\{(\mathbf{x}_i, y_i) \in \mathscr{D}_{\text{test}} : M(\mathbf{x}_i) = y_i\}|}{|\mathscr{D}_{\text{test}}|}. \tag{2.5}$$

Figure 2.1 illustrates classification with training and testing sets for gene expression data.

K-Nearest Neighbours (k-NN) The *k-nearest neighbour classifier* (k-NN) is a popular classification method were the class label of an instance $\mathbf{x}^*$ is predicted according to majority vote by its k nearest neighbour from the training set $\mathscr{D}_{\text{train}}$. The similarity measured by some distance function $d(\cdot,\cdot)$. Distance measures are discussed in detail in Section 3.1.

A crucial aspect of nearest neighbour classification is its *interpretability*. The prediction output is decided only by the k nearest neighbours of $\mathbf{x}^*$. These few vectors and the corresponding patients can be inspected by a human expert and further conclusions may be drawn.

Support Vector Machines (SVM) Support Vector Machines employ projections of the data to very high-dimensional spaces and attempt to linearily separate the classes in data with a *maximum-margin* hyperplane, which is as far from the points as possible. In practice, dual representations are used, and the projections are realised by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$.

Support Vector Machines can be successfully applied to gene expression data (W. Lin and Chen, 2013; Mukherjee, 2003). For a detailed overview of SVMs and other sparse kernel machines, we refer to Bishop (2006, Chapter 7).

3 Hubness in Gene Expression Data

Gene expression data sets contain expression values of thousands of genes. This means instances in gene expression data sets are represented by vectors in very high dimensional Euclidean space.

For example, the Breast Cancer data set (Sotiriou et al., 2003) contains 7650 features measured by cDNA microarray chips. Therefore, instances of the Breast Cancer data set are vectors in $\mathbb{R}^{7650}$, i.e. the space of 7650-dimensional vectors.

Data sets with such high dimensionality give rise to a collection of phenomena known as the *curse of dimensionality*. A prominent aspect of the 'curse' is *hubness* (Radovanovic et al., 2010a). Hubness is the emergence of *hubs* in the data set, instances which are similar to a surprisingly large number of other instances.

3.0.1 Scale-Free Netorks

The terminology *hub* comes from Albert et al. (1999) in the context of scale-free network analysis.

Scale-free networks are random graphs whose vertex degrees are distributed according to a power law,

$$P(\text{out-deg}\, x = n) \propto n^{-\gamma}.$$

Scale-free networks include

- the World Wide Web with hyperlinks between sites as edges,
- the Internet with network connections as edges,
- movie actor and scientist collaboration graphs,
- celluar chemical reaction networks
- and ecological networks.

In these graphs, there are a few vertices with suprising large number of adjacent edges. These vertices, located in the thick 'tails' of the power law distribution, are hubs of scale-free networks.

In the rest of this Chapter, we will consider networks created from gene expression data sets according to the similarity of gene expression profiles.

3.1 Distance Metrics

To determine similarity of instances in gene expression data, we consider four widely used distance measures, Euclidean, Manhattan, maximum and cosine distance.

The Euclidean and Manhattan distances are special cases of the ℓ_r or Minkowski distance:

Definition 3.1 *The ℓ_r or* Minkowski *distance ($r > 0$) of two gene expression vectors is calculated as*

$$d_r(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_r = \left(\sum_t |x_{i,t} - x_{j,t}|^r\right)^{1/r}, \tag{3.1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the gene expression profiles of the ith and jth patiens, and $x_{i,t}$ is the expression value of the ith patient and tth gene.

Definition 3.2 Euclidean distance *is the ℓ_2-distance of two vectors,*

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 = \sqrt{\sum_t (x_{i,t} - x_{j,t})^2}. \tag{3.2}$$

Definition 3.3 Manhattan distance *is the ℓ_1-distance of two vectors,*

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_1 - \mathbf{x}_j\|_1 = \sum_t |x_{i,t} - x_{j,t}|. \tag{3.3}$$

Definition 3.4 Maximum *or supremum* distance *of two vector, often denoted as their ℓ_∞ distance, is the maximum of their componentwise absolute differences,*

$$d_\infty(\mathbf{x}_i, \mathbf{x}_j) = \max_t = |x_{i,t} - x_{j,t}|. \tag{3.4}$$

Definition 3.5 Cosine similarity *is the cosine of the angle between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$,*

$$\cos\theta_{i,j} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} = \frac{\sum_t x_{i,t} x_{j,t}}{\sqrt{\sum_t x_{i,t}^2} \cdot \sqrt{\sum_t x_{j,t}^2}}. \tag{3.5}$$

Cosine distance is defined to be small if two instances a similar, thus, it can be used with distance-based algorithms such as k-nearest neighbours. It has the form

$$d_{\cos}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \cos\theta_{i,j}. \tag{3.6}$$

Note that the cosine distance is not a proper metric, because the *triangle equality*

$$d_{\cos}(\mathbf{x}_1, \mathbf{x}_2) \le d_{\cos}(\mathbf{x}_1, \mathbf{x}_3) + d_{\cos}(\mathbf{x}_2, \mathbf{x}_3)$$

may not always be satisfied. The classification algorithms used in the work do not require the distance metric to satisfy the metric axioms, therefore, cosine distance may be employed without modification.

3.2 Nearest Neighbour Graph

We can construct the directed k-nearest neighbour graph $G = (V, \vec{E})$ of a data set $\mathscr{X}$ according to a distance function $d(\cdot,\cdot)$ as follows:

1. The vertices of the nearest neighbour graph are instances of the data set $V = \mathscr{X}$. In our application, this means G has a vertex for each patient's gene expression profile.

2. G contains the directed edge $x \to x'$ between instances $x, x' \in \mathscr{X}$ if and only if x is among the k nearest neighbours of x' according to the distance $d(\cdot,\cdot)$.

In this Section, we will focus on the case where $d(\cdot,\cdot)$ is one of Euclidean, Manhattan, maximum an cosine distances.

The k-nearest neighbour graph plays an important role in instance selection (Buza et al., 2011) and semi-supervied classification (Marussy and Buza, 2013) problems.

An important property of vertices in V is their out-degree, which is the number of their outgoing edges. This is the number of other instances for which they may influence the prediction output of a k-nearest neighbour classifier.

We call the out-degree of some vertex $x \in V$ of the k-nearest neighbour graph its *k-occurrence score* $N_k(x)$. Alternatively, the following definition may be used:

	n	c	d	d_{mle}
Breast Cancer	95	2	7650	25.95
Breast Cancer 2000	2	95	2000	19.62
Colon Cancer	62	2	2000	13.22
Lung Cancer	203	5	12600	19.30
Lung Cancer 2000	203	2	2000	16.79

Table 3.1 Number of instances (n), number of classes (c), number of genes (d) and intrinsic dimensionality (d_{mle}) in the data sets we used for hubness measurements.

Definition 3.6 (adapted from Tomasev et al. (2014)) *Let $\mathcal{N}_k(\mathbf{x}') \subset \mathcal{X}$ denote the k nearest neighbours of $\mathbf{x}'$ according to a chosen distance metric $d(\cdot,\cdot)$. The k-occurrence score of an instance $\mathbf{x}$, denoted by $N_k(\mathbf{x})$, is number of other instances which have $\mathbf{x}$ among their k nearest neighbours neighbours*

$$N_k(\mathbf{x}) = |\{\mathbf{x}' \in \mathcal{X} : \mathbf{x} \in \mathcal{N}_k(\mathbf{x}') \wedge \mathbf{x} \neq \mathbf{x}'\}|. \tag{3.7}$$

Hubs in the nearest neighbour graph have large out-degree, therefore, they have surprisingly high N_k.

3.2.1 Nearest Neighbour Graphs in High Dimensions

As the dimensionality of data sets and the number of instances goes to infinity, nearest-neighbour graphs in synthethic data become similar to scale-free networks, that is, there is and emergence of hubs (Radovanovic, 2011). Therefore, hubness in as aspect of the 'curse of dimensionality'.

With Euclidean distances, behaviour is caused the asymptotic behaviour of the noncentral χ-distribution (Radovanovic et al., 2010a), which is related to the *concentration of distances* phenomenon (Aggarwal et al., 2001). For cosine distances, analogous results exist (Nanopoulos et al., 2009).

In reald world data sets, the emergence of hubness depends on the *intrinsic dimensionality* of the data set instead of its embedding dimensionality Radovanovic et al. (2010b). Therefore, hubness is and aspect of the 'curse of intrinsic dimensionality'.

Intrinsic dimensionality refers to the dimensionality of the data set regardless of its embedding. For example, a data set consisting of a 2-dimensional plane represented as d-dimensional vectors with $d > 2$ still has intrinsic dimensionality of 2.

3.2.2 Degree Distributions in Gene Expression Data

To illustrate hubness in gene expression data sets, we plot the degree distributions of the $k = 5$ nearest neighbour graphs of three gene expression data sets and two derived data sets, which are summarised in Table 3.1.

The intrinsic dimensionality of the data sets were estimated with a maximum likelihood estimator based on the Poisson distribution (Levina and Bickel, 2004). We used the same number of neighbours ($k = 5$) for estimation of intrinsic dimensionality that was used for the calculation of hubness indicators.

- The *Breast Cancer* data set (Sotiriou et al., 2003) contains expression values for 7650 genes of 95 breast cancer specimens. 32 of these specimens were estrogen receptor negative (ER–), while 63 were ER+.

  The original data set published by Sotiriou et al. contains 4 expression vectors for 4 additional specimens, for which the the ligand-binding assays and immunohistochemistry method of determining ER status gave condtradictory results. We removed these vectors from the data set, so evaluations are only performed on the instances with disambigous class label assignment.

- The *Colon Cancer* data set contains (Alon et al., 1999) 62 colon tissue sample instances, 40 of which were cancerous.

  Out of the more than 6500 genes measures, the data set contains the expression values for only the 2000 with highes minimal expression value. In other words, the the genes were sorted descending according to

  $$x_{\min,t} = \min_i x_{i,t}$$

  and only the top 2000 were preserved.

- The *Lung Cancer* data set (Bhattacharjee et al., 2001) contains 203 specimens with 12600 genes. There are 139 adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcionoids and 6 small-cell carcinoma cases among the specimens in addition to 17 normal lung samples.

- In order to distinguish the artifacts in the data sets which are caused by pre-processing, we applied the same pre-processing step to the Breast Cancer and Lung Cancer data sets that Alon et al. used for the Colon Cancer data set. The results are the *Breast Cancer 2000* and *Lung Cancer 2000* data sets with the 2000 most expressed genes from the respective data sets.

On the Breast Cancer data set, Euclidean and Manhattan distances exhibit similar k-occurrence score distributions (Figure 3.1). Over 40% of the instances are 'anti-hubs' with $N_5 = 0$ or 1. There are a few instances with k-occurrence scores up to 31 in the tail of the distribution which are hubs. Maximum and cosine distances have comparable number of hubs, albeit a smaller fraction of the data set are anti-hubs.

The pre-processing in the Breast Cancer 2000 data set does not impact hubs considerably. However, it has reduced the number of anti-hubs for Euclidean and Manhattan distances, which became not much greater than that of maximum and cosine distances.

In Table 3.1 we can see that while the pre-processing removed 74% of the attributes, the intrinsic dimensionality was decreased only by 24%. This explains why is the change in the shape of the curves relatively small.

The distributions for the Colon Cancer data set are very noisy (Figure 3.2), probably because of the relatively small size (62 instances) of the data set. We can still observe some hubs with $N_5 \geq 10$ in the tails of the distributions.

The k-occurrence score distributions of the Lung Cancer data are extremely skewed and similar to the distributions of the Breast Cancer data set (Figure 3.3).

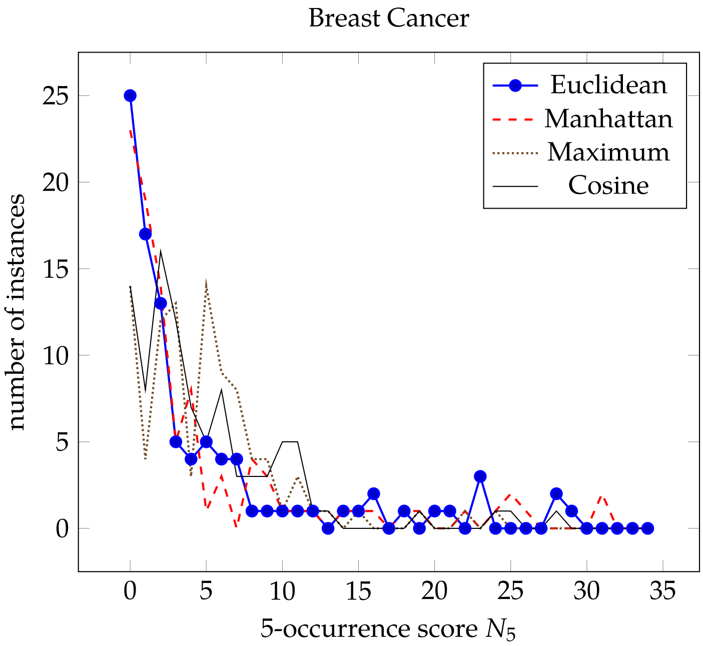
The removal of 84% of attributes resulted in a mere 13% decrease of intrinsic dimensionality and left the general shape of the distributions intact. However, there was an increase in the number of anti-hubs for Euclidean and Manhattan distances.

3.3 Measurement of Hubness

Radovanovic et al. (2010a) suggested the *skewness* (standardised third moment) of the distribution of k-occurences as an indicator of hubness,

$$\mathcal{S}_{N_k} = \frac{\mathbb{E}[(N_k(\mathbf{x}) - \mu_{N_k})^3]}{\sigma_{N_k}^3}, \tag{3.8}$$

where μ_{N_k} and σ_{N_k} are the mean and standard deviation of the distribution of N_k, respectively. When the skewness $\mathcal{S}_{N_k}$ is positive, the distribution is right-tailed. The tail of a highly skewed k-occurence distribution contains instances with hubness values much larger than the mean.



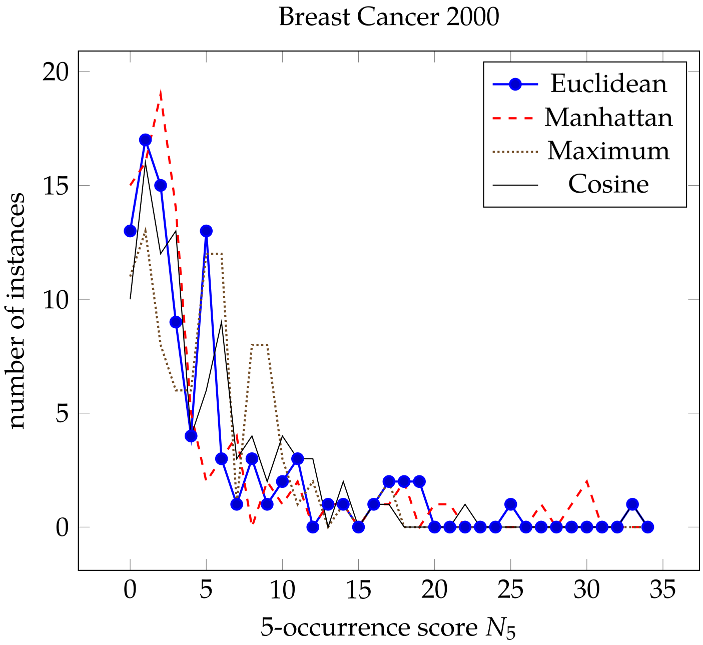


Figure 3.1 Distribution of k-occurrence scores in the Breast Cancer data sets with Euclidean, Manhattan, maximum and cosine distances and $k = 5$.

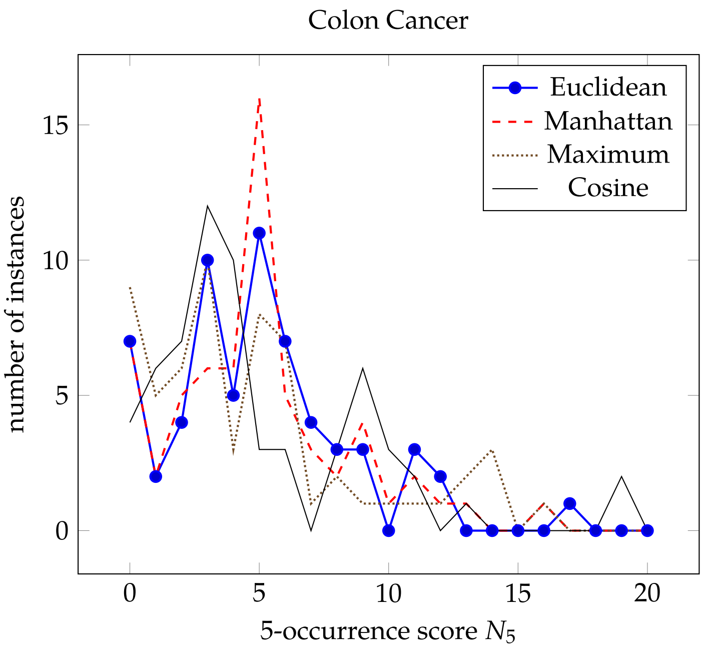


Figure 3.2 Distribution of k-occurrence scores in the Colon Cancer data set with Euclidean, Manhattan, maximum and cosine distances and $k = 5$.

Another two indicators, k-anti-hub occurrence and k-hub occurrence, were proposed by Schnitzer and Flexer (2014):

Definition 3.7 *The k-anti-hub occurrence score (A^k_{occ}) of a dataset $\mathscr{X}$ is the fraction of instances with zero k-occurrence score, i.e.*

$$A^k_{occ} = \frac{|\{\mathbf{x} : \mathbf{x} \in \mathscr{X} \land N_k(x) = 0\}|}{|\mathscr{X}|}. \tag{3.9}$$

Definition 3.8 *The k-hub occurrence score (H^k_{occ}) of a dataset $\mathscr{X}$ is normalised sum of the k-occurrence scores of instances with k-occurrence at least* $2k$*, i.e.*

$$H^k_{occ} = \frac{1}{k|\mathscr{X}|} \sum_{\substack{\mathbf{x} \in \mathscr{X} \\ N_k(\mathbf{x}) \geq 2k}} N_k(\mathbf{x}). \tag{3.10}$$

A^k_{occ} and H^k_{occ} were found to be more stable indicators of hubness in the context of distance function selection that the skewness $\mathscr{S}_{N_k}$.

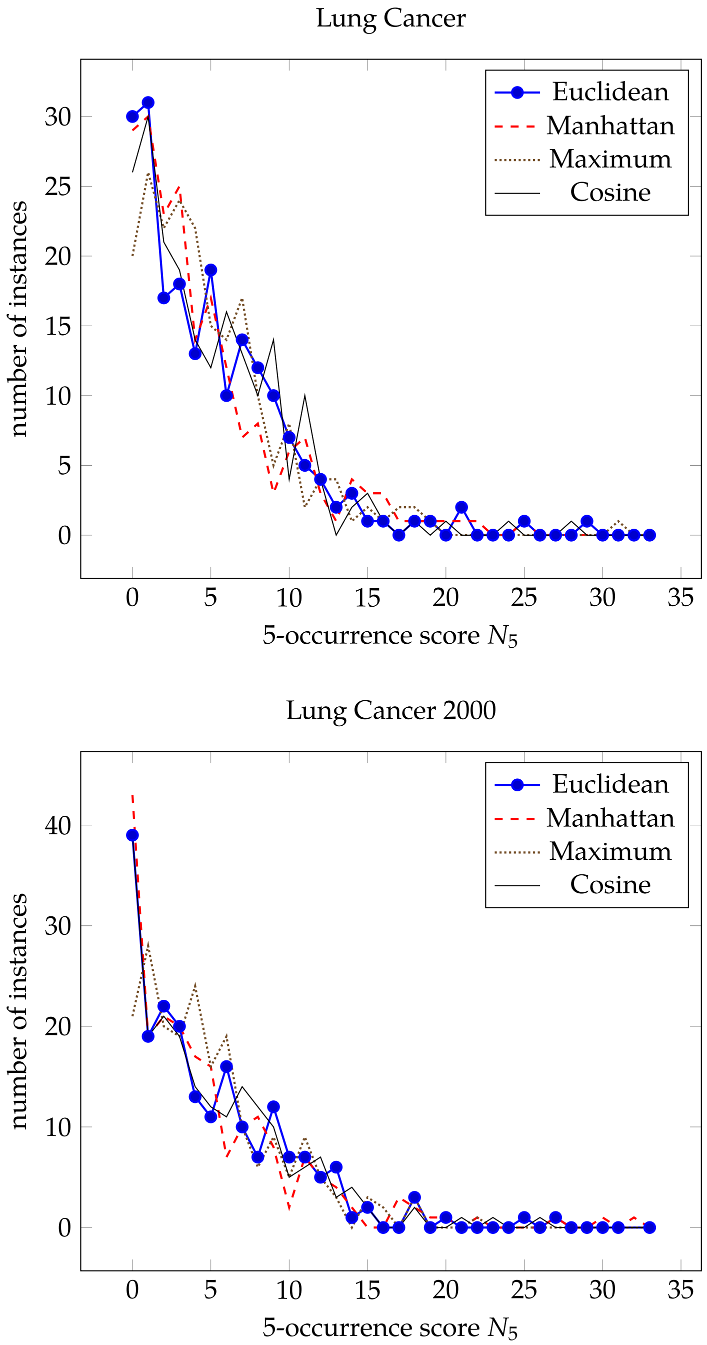
3.3.1 Bad Hubness

Bad hubness is the tendency of hubs to have different class labels than the instances they are nearest neighbour of. This phenomenon may mislead classifiers and degrade classification accuracy (Radovanovic et al., 2010b).

To study the relationship between hubs and class labels in labeled data set $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we use class-conditional k-occurrence socres:

Definition 3.9 (adapted from Tomasev et al. (2014)) *Let $\mathscr{N}_{k,y}(\mathbf{x}') \subseteq \mathscr{N}_k(\mathbf{x}')$ denote the subsetset of other instances that have* $\mathbf{x}$ *among their k nearest neighbours and belong to the class y, i.e.*

$$\mathscr{N}_{k,y}(\mathbf{x}) = \{\mathbf{x}_i : (\mathbf{x}_i, y_i) \in \mathscr{D} \land \mathbf{x}_i \in \mathscr{N}_k(\mathbf{x}) \land y_i = y\}. \tag{3.11}$$



Lung Cancer

Lung Cancer 2000

Figure 3.3 Distribution of k-occurrence scores in the Lung Cancer data sets with Euclidean, Manhattan, maximum and cosine distances and $k = 5$.

The class-conditional k-occurrence score $N_{k,y}(\mathbf{x})$ *of an instance* $\mathbf{x}$ *is the size of number of instances of class* y *that have* $\mathbf{x}$ *among their* k *nearest neighbours, i.e.*

$$N_{k,y}(\mathbf{x}) = |\{\mathbf{x}' \in \mathscr{D} : y(\mathbf{x}') = y\}|. \tag{3.12}$$

The k-occurrence score of $\mathbf{x}$ may be decomposed as a sum of class-conditional scores,

$$N_k(\mathbf{x}) = \sum_{y \in \mathscr{Y}} N_{k,y}(\mathbf{x}).$$

The most interesing hubness-related parameter for quantifying bad hubness is the *bad k-occurrence score* $BN_k(\mathbf{x})$, which is the number of other instances $\mathbf{x}'$ which have $\mathbf{x}$ among their k nearest neighbours but belong to a *different* class than x. This quantity can be calculated as

$$BN_k(\mathbf{x}) = N_k(\mathbf{x}) - N_{k,y(\mathbf{x})}(\mathbf{x}). \tag{3.13}$$

To capture the bad hubness in the whole data set, we use the normalised total bad hubness proposed by Radovanovic et al. (2010b):

Definition 3.10 *The* normalised total bad hubness $\widetilde{TBN}_k$ *in a labeled data set* $\mathscr{D}$ *is the sum of bad k-occurrence scores divided by the sum of all k-occurrences,*

$$\widetilde{TBN}_k = \frac{\sum_{\mathbf{x} \in \mathscr{D}} BN_k(\mathbf{x})}{\sum_{\mathbf{x} \in \mathscr{D}} N_k(\mathbf{x})}. \tag{3.14}$$

Each instance $\mathbf{x}'$ has exactly k other instances $\{\mathbf{x_i}\}_{i=1}^k$ with $\mathbf{x}' \in \mathscr{N}_k(\mathbf{x})$, since these $\{\mathbf{x}_i\}$ instances are the k nearest neighbours of $\mathbf{x}'$ by the definition of $\mathscr{N}_k(\mathbf{x})$. Therefore, the sum $\sum_{\mathbf{x} \in \mathscr{D}} N_k(\mathbf{x}) = k|\mathscr{D}|$ and $\widetilde{TBN}_k$ may be written as

$$\widetilde{TBN}_k = \frac{1}{k|\mathscr{D}|} \sum_{\mathbf{x} \in \mathscr{D}} BN_k(\mathbf{x}). \tag{3.15}$$

3.3.2 Measurements in Gene Expression Data Sets

We calculated the skewness of 5-occurrences $\mathscr{S}_{N_k}$, the 5-anti-hub occurrence score A^5_{occ}, the k-hub occurrence score H^5_{occ} and the normalised total bad hubness $\widetilde{TBN}_5$ of the Breast Cancer, Breast Cancer 2000 and Colon Cancer data sets. Table 3.2 displays the results for Euclidean, Manhattan, maximum and cosine distances.

To study the effects of dimensionality reduction, we have extracted the first 20 principal compoments of the data sets with the `stats` R package (R Core Team, 2014) and calculate the hubness indicators for the low-dimensional representations, too. The number of principal components was selected to be approximately equal to the intrinsic dimensionality d_{mle} of the data sets (Table 3.1).

With the original representation, the distributions of k-occurrence scores is highly skewed in all data sets with all distance functions, as we have seen in Section . Surprisingly, the multi-class Lung Cancer data set

The extraction of principal components greatly decreases the indicators of hubness $\mathscr{S}_{N_k}$ and A^5_{occ}, while slightly decreases H^5_{occ}. However, the total bad hubness $\widetilde{TBN}_k$ remains similar to the original representation's values. This agrees with previous result the dimensionaly reduction cannot mitigate bad hubness (Radovanovic, 2011), unless the number of resulting dimensions is less than the intrinsic dimensionality of the data set. However, such aggressive dimensionality reduction leads to significant loss of information.

	Originial Representation				20 Principal Components			
	$\mathcal{S}_{N_5}$	A^5_{occ}	H^5_{occ}	$\widetilde{TBN}_5$	$\mathcal{S}_{N_5}$	A^5_{occ}	H^5_{occ}	$\widetilde{TBN}_5$
Breast Cancer								
Euclidean	1.889	0.263	0.646	0.246	1.334	0.084	0.406	0.206
Manhattan	2.028	0.242	0.629	0.267	1.281	0.105	0.385	0.240
Maximum	1.768	0.147	0.332	0.263	0.858	0.116	0.371	0.183
Cosine	2.101	0.147	0.476	0.238	0.503	0.032	0.152	0.227
Breast Cancer 2000								
Euclidean	2.105	0.137	0.552	0.225	1.174	0.842	0.339	0.223
Manhattan	2.194	0.158	0.600	0.259	1.296	0.105	0.406	0.248
Maximum	0.850	0.116	0.272	0.373	1.041	0.632	0.250	0.234
Cosine	2.270	0.105	0.474	0.270	0.471	0.021	0.160	0.257
Colon Cancer								
Euclidean	0.859	0.113	0.239	0.316	0.998	0.097	0.268	0.306
Manhattan	0.783	0.113	0.235	0.316	0.866	0.065	0.320	0.310
Maximum	1.220	0.145	0.442	0.374	0.929	0.065	0.210	0.330
Cosine	1.419	0.065	0.332	0.345	1.227	0.065	0.252	0.335
Lung Cancer								
Euclidean	1.665	0.148	0.339	0.096	0.862	0.054	0.286	0.096
Manhattan	1.908	0.142	0.503	0.106	1.051	0.049	0.261	0.096
Maximum	1.778	0.099	0.376	0.193	0.985	0.034	0.193	0.111
Cosine	1.499	0.128	0.371	0.100	0.541	0.034	0.201	0.117
Lung Cancer 2000								
Euclidean	1.387	0.192	0.448	0.115	0.984	0.069	0.347	0.107
Manhattan	1.948	0.212	0.467	0.103	1.063	0.059	0.342	0.112
Maximum	1.167	0.103	0.395	0.197	1.092	0.044	0.299	0.130
Cosine	1.285	0.192	0.425	0.110	0.488	0.025	0.141	0.126

Table 3.2 Hubness measurements in the considered data sets with the original gene expression vector representations and with the first 20 principal components extracted.

4 Hubness-aware Gene Expression Classification

In the previous Chapter, we demonstrated the prominent presence of hubs and bad hubs in gene expression data sets. Because hubness, especially bad hubness can significantly deteriorate classification accuracy, we expect classification algorithms which were designed with hubness in mind to outperform standard algorithms.

In this Chapter, we report the findings of our empirical evaluation of stadard and hubness-aware classifiers on gene expression data.

4.1 Hubness-aware Classifiers

Hubness-aware classifiers are modifications of the nearest-neighbour classification paradigm that operate under the assumption of hubness. They were designed to take advantage of the presence of hubness either explicitly or implicitly and mitigate the bad hubness artifacts. They outperform standard classifiers in various domains such as text and image (Tomasev and Mladenic, 2012) and time-series classification (Tomasev et al., 2014).

In this Section, we present the hubness-aware classifiers and evaluate their performance on gene expression data sets by comparing them to state-of-the-art classifiers as baseline.

In the following overview of hubness-aware classifiers, we assume that the test set $\mathscr{D}_{\text{test}}$ is not available to the classifier at learning time. Thus, the hubness indicator values $N_k(\mathbf{x})$, $N_{k,y}(\mathbf{x})$ and $BN_k(\mathbf{x})$ from Chapter 3 are only calculated from the training set $\mathscr{D}_{\text{train}}$.

For a more in-depth description of hubness-aware classifiers including many examples, we refer to Tomasev and Mladenic (2012) and our recent survey Tomasev et al. (2014).

4.1.1 Hwknn: Hubness Weighted k-Nearest Neighbours

The earliest algorithm for hubness-aware classification was suggested by Radovanovic et al. (2009). Hwknn modifies k-nearest neighbour classification by weighting the votes of each neighbour by

$$h_b(\mathbf{x}_i) = \frac{BN_k(\mathbf{x}_i) - \mu_{BN_k}}{\sigma_{BN_k}}, \tag{4.1}$$

where μ_{BN_k} and σ_{BN_k} are the mean and standard deviation of BN_k, respectively. Intuitively, less weight is given to instances with high bad hubness scores.

4.1.2 Hfnn: Hubness-based Fuzzy Nearest Neighbour

Hubness-based Fuzzy Nearest Neighbour (Tomasev et al., 2011b) interprets the relative class hubness

$$u_y(\mathbf{x}) = \frac{N_{k,y}(\mathbf{x})}{N_k(\mathbf{x})} \tag{4.2}$$

as the fuzzyness of the event that **x** occurs as one of the neighbours of the point being classified. The probability of the instance $\mathbf{x}^*$ to belong to class y hence be estimated as

$$u_y(\mathbf{x}^*) = \frac{\sum_{\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)} u_y(\mathbf{x})}{\sum_{\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)} \sum_{y' \in \mathcal{Y}} u_{y'}(\mathbf{x})}. \tag{4.3}$$

4.1.3 Hnbnn: Naive Hubness Bayesian k-Nearest Neighbour

Each k-occurrence can be treated as a random event. Hnbnn performs a Naive-Bayesian inference based on these k events (Tomasev et al., 2011a),

$$P(y* = y | \mathcal{N}_k(\mathbf{x}^*)) \propto P(y) \prod_{\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)} P(\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*) | y). \tag{4.4}$$

$P(y)$ denotes the probabilty that an instance belongs to class y and may be estimated as

$$P(y) = \frac{|\mathcal{D}_{\text{train},y}|}{|\mathcal{D}_{\text{train}}|}, \tag{4.5}$$

where $\mathcal{D}_{\text{train},y}$ is the set of instances from the training set that belong to class y.

$P(\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)|y)$ denotes the probability that **x** appears among the k nearest neighbours of and instance from class y, therefore, it can be estimated as

$$P(\mathbf{x} \in \mathcal{N}_k(\mathbf{x}^*)|y) = \frac{|\mathcal{N}_{k,y}(\mathbf{x})|}{|\mathcal{D}_{\text{train},y}|}. \tag{4.6}$$

4.1.4 Hiknn: Hubness Information k-Nearest Neighbour

Hubness Information k-Nearest Neighbour uses the self-information $I(\mathbf{x})$ associated with the event that **x** appears as a nearest neighbour

$$I(\mathbf{x}) = \log \frac{1}{P(\mathbf{x} \in \mathcal{N}_k)} \qquad P(\mathbf{x} \in \mathcal{N}_k) \approx \frac{N_k(\mathbf{x})}{|\mathcal{D}_{\text{train}}|} \tag{4.7}$$

to determine relative and absolute relevance factors of hubs to aid classification (Tomasev and Mladenic, 2012).

4.2 Evaluation Methods

4.2.1 Data Sets

We conduct the empirical evaluation of the Breast Cancer, Colon Cancer and Lung Cancer data sets presented in Section 3.3.2 and Table 3.1.

The classifiers are run on both the original, high dimensional representations of gene expression vectors and the 20-dimensional representations obtained by Principac Component Analysis. However, we do not use the derived Breast Cancer 2000 and Lung Cancer 2000 data sets, which were generated merely by discarding attributes.

4.2.2 Baselines

We consider the following baselines in out evaluations:

1. The *k-nearest neighbour classifier* (k-NN) with either Euclidean, Manhattan, maximum and cosine distances,

2. Support Vector Machines (SVMs) with inhomogenous polynomial kernels, i.e.

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d. \tag{4.8}$$

4.2.3 Cross-Validation and Hyperparameter Search

We use 10 × 10-fold cross-validation to evaluate the classifiers, that is, the instances are randomly split into 10 folds 10 times. Each fold in each split takes the role of the test set $\mathscr{D}_{\text{test}}$ once, while the union of the remaining splits is the training set $\mathscr{D}_{\text{train}}$. This amounts to 100 runs per classifier per data set.

The classifier hyperparameters are learned with grid search and inner 10-fold cross-validation on the traing data only. The train set $\mathscr{D}_{\text{train}}$ is split into 10 folds and average classification accuracy is evaluated for each setting of hyperparameters. The final classifier learns from the whole training set $\mathscr{D}_{\text{train}}$ with the best performing hyperparameter setting.

The SVM classifiers' cost hyperparameter is searched in the range $C = 10^{-5}, 10^{-4}, \ldots, 10^{5}$. The exponent of the inhomogenous polynomial kernel d is searched in the range $d = 1, 2, 3$. That is, the parameters are grid searched from

$$(C, d) = (10^{-5}, 1), (10^{-5}, 2), (10^{-5}, 3), (10^{-4}, 1), \ldots, (10^{5}, 3).$$

For the k-NN and hubness-aware classifiers, both the distance function and the number of neighbours k is treated as a hyperparameter. The distance function is selected from Euclidean, Manhattan, maximum and cosine distances. For k-nearest neighbours, HWKNN, HFNN and NHBNN, the number of neighbours is selected from $k = 1, 2, \ldots, 10$. For HIKNN, the range of k is reduced to $k = 4, 5, \ldots, 10$ in order to avoid singularities in the information claculation.

4.2.4 Implementation

Due to the extensive hyperparameter search and cross-validation, the experiments took several days of CPU time in total on a second-generation Intel®Core$^{\text{TM}}$ i7 machine with 8 GB of RAM.

The experimental framework was implemented in a combination of C++ and Java, which communicate through the JNI Invocation API (Oracle, 2011). The hubness-aware classifiers were implented in Java by Krisztián Buza[1] and their source code is available on request.

We use the LIBSVM library (Chang and C. Lin, 2011) for Support Vector Classification and the `stats` R package (R Core Team, 2014) for Principal Component Analysis.

4.3 Results and Discussion

We report the accuracy of all studied classifiers on both the Breast Cancer and Colon Cancer data set averaged over 10 × 10 folds in Table 4.1 and Table 4.2. Statistical significance of results was evaluated with two-tailed paired difference t-test at significance level $p < 0.05$.

Surprisingly, the hubness-aware classification algorithms were not only less accurate that Support Vector Machines, they were also significantly outperformed in many cases by k-nearest neighbour. This result is very interesting, because hubness-aware classifiers were designed to avoid the bad hubness artifacts that affects nearest-neighbour classification.

Hubness-aware classifiers performed the worst on the Colon Cancer, which was the smallest of the studied data sets with only 62 instances. Estimates of hubness in $\mathscr{D}_{\text{train}} \cup \mathscr{D}_{\text{test}}$ calculated from only $\mathscr{D}_{\text{train}}$ may be unreliable in such small data sets.

The small training set can also hurt hyperparameter search. Figure 4.1 displays the values of the neighbourhood size hyperparameter k of HWKNN selected by inner-cross validation over 10 × 10-fold cross validation. It is apparent the selection of k is least concentrated in the Colon Cancer data set. In contrast, for the Lung Cancer data set, which consists of 203 instances, the grid search selected $k = 3$ in nearly half of the folds.

The supremacy of SVMs in out experiments can be explained by fact that, being a maximum-margin classifier and depending on only a small number of support vectors,

[1] chrisbuza@yahoo.com

Original Representaion	Baseline		Hubness-aware			
	k-NN	SVM	Hwknn	Hfnn	Nhbnn	Hiknn
Breast Cancer	82.67%	**86.67%**	83.44%	84.44%	85.00%	80.56%
Significance vs *k*-NN	n/a	○	–	–	○	–
Significance vs SVM	●	n/a	●	●	–	●
Colon Cancer	74.50%	**89.33%**	71.67%	67.00%	68.83%	70.83%
Significance vs *k*-NN	n/a	○	–	●	●	●
Significance vs SVM	●	n/a	●	●	●	●
Lung Cancer	92.55%	**93.45%**	93.35%	91.80%	80.65%	91.50%
Significance vs *k*-NN	n/a	–	–	–	●	–
Significance vs SVM	–	n/a	–	●	●	●

Table 4.1 Accuracy of studied classifiers and the statistical significance of their differences. The best performing classifier is highlighed in **bold**. Significantly ($p < 0.05$) better accuracy is denoted by ○, while significantly worse accuracy is denoted by ●.

20 Principal Components	Baseline		Hubness-aware			
	k-NN	SVM	Hwknn	Hfnn	Nhbnn	Hiknn
Breast Cancer	84.22%	**88.22%**	81.44%	82.78%	82.67%	83.44%
Significance vs *k*-NN	n/a,–	○,○	●,–	–,–	–,–	–,–
Significance vs SVM	●,–	n/a,○	●,●	●,●	●,●	●,●
Colon Cancer	75.33%	**79.67%**	73.00%	67.33%	65.17%	70.17%
Significance vs *k*-NN	n/a,–	○,○	–,–	●,●	●,●	●,●
Significance vs SVM	●,●	n/a,●	●,●	●,●	●,●	●,●
Lung Cancer	91.35%	**92.45%**	91.60%	90.90%	81.60%	91.20%
Significance vs *k*-NN	n/a,–	–,–	–,–	–,●	●,●	–,–
Significance vs SVM	–,●	n/a,–	–,●	●,●	●,●	●,●

Table 4.2 Accuracy of studied classifiers and the statistical significance of their differences after unsupervised extraction of the first 20 principal components for the data. The first symbol refers to statistical significance compared to the baseline learned with principal components extraction, while the second symbol refers to significance compared to the baseline with the originial gene representation.

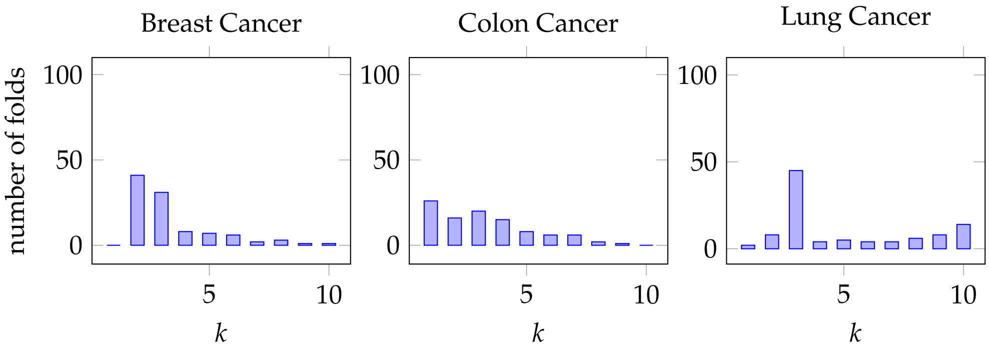


Figure 4.1 Values of the neighbourhood size hyperparameter k learned by Hwknn classifier.

Support Vector Machines are suitable for classification even in very sparse spaces (Mukherjee, 2003). Moreover, bad hubs can often be good support vectors (Radovanovic et al., 2009), which makes SVMs even more suitable for gene expression data.

5 Conclusions and Future Work

In this paper, we have studied the emergence of hubness and hubness-aware classification in three real-world gene expression data sets from cancer research. Our contributions include:

- The measurement of hubness in the studies gene expression data set with Euclidean, Manhattan, maximum and cosine distances in Chapter 3.

  We have demonstrated that hubness occurs with all four distances metrics and is not an artifact of pre-processing. Additionally, we have shown that bad hubness cannot be entirely removed by dimensionality reduction via Principal Component Analysis.

  More abstractly, we can interpret the results that one should expect a degree of hubness in any gene expression data set.

- The suprising result in Chapter 4 that, despite the existence of bad hubs in gene expression data sets, state-of-the-art hubness-aware classifiers cannot outperform the simple k-nearest neighbour classifier.

  Support Vector Machine were able to cope with the small size of the data sets much better than instance-based nearest-neighbour and hubness-aware classifiers, which were plagued by unstable hyperparameter estimates.

The current hubness-aware classification algorithm seem to be insufficiently roboust for the 'small n, large p' challange that gene expression data pose. The development of more roboust hubness measurements and hubness-aware classifiers for gene expression data seem promising and interesing research directions.

Acknowledgement This work was performed within the framework of the Hungarian Scientific Research Fund (grant no. OTKA 111710). This research was partially funded by the National Brain Research Program (project no. KTIA_13_NAP-A-III/6, project id KTIA_NAP_13-1-2013-0001).

References

Aggarwal, Charu C., Alexander Hinneburg and Daniel A. Keim (2001). 'On the Surprising Behavior of Distance Metrics in High Dimensional Spaces'. In: *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings.* Vol. 1973, pp. 420–434.

Albert, Réka, Hawoong Jeong and Albert-László Barabási (1999). 'The diameter of the world wide web'. In: *CoRR* cond-mat/9907038.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine (1999). 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays'. In: *Proc. Natl. Acad. Sci. U.S.A.* 96.12. Data set publicly available at `http://genomics-pubs.princeton.edu/oncology/`, pp. 6745–6750.

Alter, O., P. O. Brown and D. Botstein (2000). 'Singular value decomposition for genome-wide expression data processing and modeling'. In: *Proc. Natl. Acad. Sci. U.S.A.* 97.18, pp. 10101–10106.

Antal, Péter, Ádám Arany, Bence Bolgár, András Gézsi, Gergely Hajós, Gábor Hullám, Péter Marx, András Millinghoffer, László Poppe and Péter Sárközy (2014). *Bioinformatics*.

Aruliah, Dhavide, Guangzhe Fan, Roderick Melnik, Suzanne Shontz, Steven Wang and Jiaping Zhu (2006). 'Nonlinear Dimension Reduction for Microarray Data (Small n and Large p)'. In: *Proceedings of the Fields–MITACS Industrial Problems Workshop*.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium'. In: *Nat. Genet.* 25.1, pp. 25–29.

Bellman, Richard Ernest (1957). *Dynamic Programming*. Princeton University Press.

Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker and M. Meyerson (2001). 'Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses'. In: *Proc. Natl. Acad. Sci. U.S.A.* 98.24. Data set publicly available at `https://www.broadinstitute.org/MPR/lung/`, pp. 13790–13795.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Buza, Krisztian, Alexandros Nanopoulos and Lars Schmidt-Thieme (2011). 'INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification'. In: *Advances in Knowledge Discovery and Data Mining - 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part II*. Vol. 6635, pp. 149–160.

Chang, Chih-Chung and Chih-Jen Lin (2011). 'LIBSVM: A library for support vector machines'. In: *ACM TIST* 2.3, p. 27.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander (1999). 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring'. In: *Science* 286.5439, pp. 531–537.

International Human Genome Sequencing Consortium (2004). 'Finishing the euchromatic sequence of the human genome'. In: *Nature* 431.7011, pp. 931–945.

Kaminski, N. and N. Friedman (2002). 'Practical approaches to analyzing results of microarray experiments'. In: *Am. J. Respir. Cell Mol. Biol.* 27.2, pp. 125–132.

Lesk, Arthur M (2012). *Introduction to Genomics, Second Edition*.

Levina, Elizaveta and Peter J. Bickel (2004). 'Maximum Likelihood Estimation of Intrinsic Dimension'. In: *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*.

Lin, Wei-Jiun and James J. Chen (2013). 'Class-imbalanced classifiers for high-dimensional data'. In: *Briefings in Bioinformatics* 14.1, pp. 13–26.

Marussy, Kristóf and Krisztian Buza (2013). 'SUCCESS: A New Approach for Semi-supervised Classification of Time-Series'. In: *Artificial Intelligence and Soft Computing - 12th International Conference, ICAISC 2013, Zakopane, Poland, June 9-13, 2013, Proceedings, Part I*. Vol. 7894, pp. 437–447.

Mukherjee, Sayan (2003). 'Classifying microarray data using support vector machines'. In: *A practical approach to microarray data analysis*, pp. 166–185.

Murthy, Sreerama K. (1998). 'Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey'. In: *Data Min. Knowl. Discov.* 2.4, pp. 345–389.

Nanopoulos, Alexandros, Milos Radovanovic and Mirjana Ivanovic (2009). 'How does high dimensionality affect collaborative filtering?' In: *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009*, pp. 293–296.

Oracle (2011). *The JNI Invocation API*. Page accessed 22nd October 2014. URL: http://docs.oracle.com/javase/7/docs/technotes/guides/jni/spec/invocation.html.

Pearson, H. (2006). 'Genetics: what is a gene?' In: *Nature* 441.7092, pp. 398–401.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Radovanovic, Milos (2011). 'Representations and Metrics in High-Dimensional Data Mining'. PhD thesis.

Radovanovic, Milos, Alexandros Nanopoulos and Mirjana Ivanovic (2009). 'Nearest neighbors in high-dimensional data: the emergence and influence of hubs'. In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*. Vol. 382, pp. 865–872.

— (2010a). 'Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data'. In: *Journal of Machine Learning Research* 11, pp. 2487–2531.

— (2010b). 'Time-Series Classification in Many Intrinsic Dimensions'. In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA*, pp. 677–688.

Schnitzer, Dominik and Arthur Flexer (2014). 'Choosing the Metric in High-Dimensional Spaces Based on Hub Analysis'. In: *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014*.

Science Exchange (2014). *Affymetrix RNA microarray lab offerings*. Page accessed 22nd October 2014. URL: https://www.scienceexchange.com/services/affymetrix-rna-microarray.

Sotiriou, C., S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris and E. T. Liu (2003). 'Breast cancer classification and prognosis based on gene expression profiles from a population-based study'. In: *Proc. Natl. Acad. Sci. U.S.A.* 100.18. Data set publicly available at http://www.pnas.org/, pp. 10393–10398.

Su, Yang, T. M. Murali, Vladimir Pavlovic, Michael Schaffer and Simon Kasif (2003). 'RankGene: identification of diagnostic genes based on expression data'. In: *Bioinformatics* 19.12, pp. 1578–1579.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles'. In: *Proc. Natl. Acad. Sci. U.S.A.* 102.43, pp. 15545–15550.

Tomasev, Nenad, Krisztian Buza, Kristóf Marussy and Piroska B. Kis (2014). 'Hubness-aware Classification, Instance Selection and Feature Construction: Survey and Extensions to Time-Series'. In: *Feature selection for data and pattern recognition (tentative title)*. Ed. by U. Stanczyk and L. Jain. To appear. Springer-Verlag.

Tomasev, Nenad and Dunja Mladenic (2012). 'Nearest neighbor voting in high dimensional data: Learning from past occurrences'. In: *Comput. Sci. Inf. Syst.* 9.2, pp. 691–712.

Tomasev, Nenad, Milos Radovanovic, Dunja Mladenic and Mirjana Ivanovic (2011a). 'A probabilistic approach to nearest-neighbor classification: naive hubness bayesian kNN'. In: *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pp. 2173–2176.

— (2011b). 'Hubness-Based Fuzzy Measures for High-Dimensional k-Nearest Neighbor Classification'. In: *Machine Learning and Data Mining in Pattern Recognition - 7th International Conference, MLDM 2011, New York, NY, USA, August 30 - September 3, 2011. Proceedings*. Vol. 6871, pp. 16–30.

Tusher, V. G., R. Tibshirani and G. Chu (2001). 'Significance analysis of microarrays applied to the ionizing radiation response'. In: *Proc. Natl. Acad. Sci. U.S.A.* 98.9, pp. 5116–5121.

Witten, Ian H., Eibe Frank and Mark A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. Morgan Kaufmann Publishers.